The Land of Disenchantment: Bias in New Mexico Teacher Evaluation Measures

by

Tray Geiger

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved January 2020 by the
Graduate Supervisory Committee:

Audrey Amrein-Beardsley, Chair
Kate Anderson
Keon McGuire
Jessica Holloway

ARIZONA STATE UNIVERSITY

May 2020

ABSTRACT

Over the past 20 years in the United States (U.S.), teachers have seen a marked shift in how teacher evaluation policies govern the evaluation of their performance. Spurred by federal mandates, teachers have been increasingly held accountable for their students' academic achievement, most notably through the use of value-added models (VAMs)—a statistically complex tool that aims to isolate and then quantify the effect of teachers on their students' achievement. This increased focus on accountability ultimately resulted in numerous lawsuits across the U.S. where teachers protested what they felt were unfair evaluations informed by invalid, unreliable, and biased measures—most notably VAMs.

While New Mexico's teacher evaluation system was labeled as a "gold standard" due to its purported ability to objectively and accurately differentiate between effective and ineffective teachers, in 2015, teachers filed suit contesting the fairness and accuracy of their evaluations. Amrein-Beardsley and Geiger's (revise and resubmit) initial analyses of the state's teacher evaluation data revealed that the four individual measures comprising teachers' overall evaluation scores showed evidence of bias, and specifically, teachers who taught in schools with different student body compositions (e.g., special education students, poorer students, gifted students) had significantly different scores than their peers. The purpose of this study was to expand upon these prior analyses by investigating whether those conclusions still held true when controlling for a variety of confounding factors at the school, class, and teacher levels, as such covariates were not included in prior analyses.

Results from multiple linear regression analyses indicated that, overall, the measures used to inform New Mexico teachers’ overall evaluation scores still showed evidence of bias by school-level student demographic factors, with VAMs potentially being the most susceptible and classroom observations being the least. This study is especially unique given the juxtaposition of such a highly touted evaluation system also being one where teachers contested its constitutionality. Study findings are important for all education stakeholders to consider, especially as teacher evaluation systems and related policies continue to be transformed.

DEDICATION

This dissertation is dedicated to my late grandma, Grandma Jo. Without you, I never would have made it.

ACKNOWLEDGEMENTS

Completing this dissertation has been my most challenging intellectual, mental, and emotional endeavor to date. While it sounds clichéd, I mean it when I say that this achievement would not have been possible without the unyielding support from many mentors, friends, colleagues, and classmates.

First and foremost, I wish to extend my heartfelt appreciation to my committee members. To Dr. Audrey Amrein-Beardsley, your support, encouragement, and generosity that began the first day we met has never wavered, even as my research trajectory, career ambitions, and personal goals shifted many times. Your efforts have shaped me into the researcher I am today, and the opportunities I received while under your wing have been transformational. While I am incredibly proud of myself for completing this dissertation, I am just as proud to be another one of your successful students.

To Dr. Kate Anderson and Dr. Keon McGuire, while our times together have been short, I am very grateful for all of our conversations that provoked me to think about the world in a completely different light. Additionally, your understandings of work-life balance allowed me to give myself grace when I often silently criticized my need for self-care, and that was more influential than you will ever know.

To Dr. Jessica Holloway, your support, compassion, respect, and friendship extends far beyond words. Under your guidance, I gained a vast array of new knowledge, including the ability to critically think about why we believe the things we do. No matter

what I was going through, you were there with open arms and ears, even from literally across the world. On many days, that made all the difference.

There are many others within the Mary Lou Fulton Teachers College to whom I feel indebted. Specifically to Dr. Margarita Pivovarova and Dr. Jeanne Powers, thank you for your mentorship, kindness, and support through the years—even when my research trajectory changed, and more than once. To Dr. Sherman Dorn, your support from behind the scenes and kind smile was always appreciated and never lost on me. I am also grateful for the generous financial and administrative support I received from MLFTC, as well as the ASU Graduate College and the Graduate and Professional Student Association at ASU.

I would also like to acknowledge Dr. Lisa McIntyre, Dr. Clarin Collins, and Dr. Rachael Gabriel. You each served as a sounding board at various times throughout this process, and also provided me with empathy and humor exactly when I needed it the most. I would like to acknowledge my colleagues, friends, and family—you all are the ones who have been the backbones of this journey. To my former and current coworkers at ASU and to my friends in the MLFTC Ph.D. programs, especially Dr. Katy Chapman and Dr. Catharyn Shelton, thank you for joining me in celebrating my wins and providing space to voice my anxieties and struggles. To Tia Norris, J.D. and the rest of the FitPro crew, especially Jen Durham, Cat Thomas, and Eva Jannotta, thank you for all of the emotional support and much-needed escapes from the grind of academia. To Dr. Alex Geiger, although our lives look drastically different than imagined, I would be remiss if I did not acknowledge the integral role you played in this journey. And finally, to my brother and his wife, Frank and Alexia Geiger, thank you for cheering me on over the past few years from across the country.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

APPENDIX Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Generally speaking, one of the main purposes of formal education and schooling is to provide students with increased knowledge and skills (Goodlad, 1979). While students learn from a variety of actors (e.g., parents, teachers, peers) and in a variety of settings (e.g., home, school, community), teachers are the within-school actors who have been tasked with providing students such knowledge and skills (Leu, 2005). To no surprise, researchers have thus found that teachers are the most impactful in-school factor that contributes to students' learning and achievement (Coleman et al., 1966). There is ample evidence that indicates that more effective teachers have larger impacts on their students' learning and achievement compared to less effective teachers (e.g., Darling-Hammond, 2000; Hanushek & Rivkin, 2010), and policymakers believe that improving teacher quality is necessary for the United States (U.S.) to protect the quality of its workforce and be globally competitive (Cochran-Smith, 2008; Furlong, Cochran-Smith, & Brennan, 2009). As such, evaluating teachers' performances and gauging their effectiveness is necessary.

While nobody disagrees that teachers, just like any other employee in any other profession, should be evaluated or have their performance assessed, there are major discrepancies about what it means to be an effective teacher (Danielson, 2016). Some identify an effective teacher merely as one who can raise students' academic achievement levels (e.g., standardized test scores; Creemers & Kyriakides, 2008; Hanushek & Rivkin, 2010; Kimball, White, Milanowski, & Borman, 2004), while others see an effective teacher as one who can not only increase students' knowledge, but can advance students' socioemotional skills (e.g., interpersonal relationships, perseverance, self-awareness; Blazar & Kraft, 2015; Jackson, 2012, 2019). Teacher evaluation is important not just for formative purposes to help teachers improve upon their instruction, but also for accountability purposes to help inform personnel decisions (Grissom & Youngs, 2016; Kennedy, 2010b; Papay, 2012). However, when it comes to teacher evaluation policy in the U.S., there is also no clear consensus about which of these two purposes, if either, should inform policy more than the other (Firestone, 2014).

Given the various ways to define effective teaching and different opinions on the purposes of teacher evaluation, teachers are typically evaluated by a variety of different tools or measures. All of these measures aim to assess the underlying construct of "teacher effectiveness," though each tool is used for slightly different purposes. To date, teachers have typically been evaluated by two measures: value-added models (VAMs) and classroom observations. VAMs, which are described in more detail below, aim to isolate and then quantify the effect that a teacher has on his or her student's standardized test scores. Classroom observations, on the other hand, are less outcomes-based and focus on a teacher's actual pedagogical and related practices (e.g., delivering instruction, managing a classroom, being prepared for lessons, professional demeanor). In the past five years, an additional measure that has become popular is the student perception survey (SPS), which allow students to rate their teachers on a variety of characteristics. These multiple measures used to assess different aspects of the teaching effectiveness construct are often combined to form one overall rating of teacher effectiveness, which is then often used for formative purposes, summative purposes, or both.

Background

With the rise of today's neoliberal thinking flourishing in the latter half of the 20th century through the early 2000s (Harvey, 2007; Hursh, 2001), along with claims that the American education system was in need of dire reform as it was purportedly failing the nation's children (e.g., National Commission on Excellence in Education [NCEE], 1983), teacher evaluation systems that were mostly or solely formative in nature fell under heavy criticism (e.g., Harris, 2011; Tucker & Stronge, 2005). The popular belief that took hold among policymakers and the American public was that teachers who were not being held accountable for contributing to their students' learning were contributing to students' declining test scores, poor school quality, and, subsequently, a weakening American workforce—all of which would result in the U.S. losing its global prominence (NCEE, 1983).

Despite many scholars' skepticisms about this purported "failing" of the country's education system (e.g., Berliner & Biddle, 1995; see also Guthrie & Springer, 2004), the discourse surrounding public education became laden with negative rhetoric about failing public schools, ineffective teachers, and "broken" teacher evaluation systems (Darling-Hammond, 2013; Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984). The result of this fear-based rhetoric led to some of the strictest federal policies (e.g., No Child Left Behind [NCLB], 2001) to hold teachers (as well as principals and schools) accountable for their students' performance and academic achievement.

The “Ideal” Distribution of Teacher Quality

In 2009, the New Teacher Project published a highly influential report, *The Widget Effect* (Weisberg, Sexton, Mulhearn, & Keeling, 2009), which highlighted how the majority of teachers across multiple districts in multiple states were classified as “effective” or better, with a paucity of teachers being classified as “ineffective.” Per Weisberg et al., this skewed distribution of teacher effectiveness was incredibly problematic and illogical given that American students were only performing on par with, at best, students from other similar industrialized nations. Weisberg et al. could not understand how so many U.S. teachers were rated so highly yet the distribution of country’s student achievement (i.e., test scores) was not similarly skewed. *The Widget Effect* report was highly publicized by the media, and Weisberg et al.’s sentiment around the seemingly illogical distribution of teacher effectiveness was echoed by numerous policymakers, as well as pockets of researchers and practitioners (e.g., Burgess, 2017; Doherty & Jacobs, 2015; Walsh, Joseph, Lakis, & Lubell, 2017).

Although *The Widget Effect*’s conclusions were not agreed upon by all (e.g., Amrein-Beardsley, 2017; Pecheone & Wei, 2009), its findings and the push for increased teacher accountability was used as evidence as to why states needed to completely reform their teacher evaluation systems. Subsequently, and in line with the multi-billion-dollar Race to the Top (RTTT) initiative that fiscally incentivized states to hold their teachers accountable (U.S. Department of Education [USDOE], 2009a), many states tried to revamp their teacher evaluation systems so as to achieve more normal distributions of teacher effectiveness. In states that classified teachers into one of five different effectiveness ratings, a more normal distribution of teacher quality would be where the majority of teachers were rated as average (i.e., "effective") with fewer yet symmetrical proportions of teachers rated as below average (i.e., "minimally effective") and above average (i.e., "highly effective"), respectively, and even fewer and symmetrical proportions of teachers rated as much below average (i.e., "ineffective) and much above average (i.e., "exemplary"), respectively (see Figure 1).

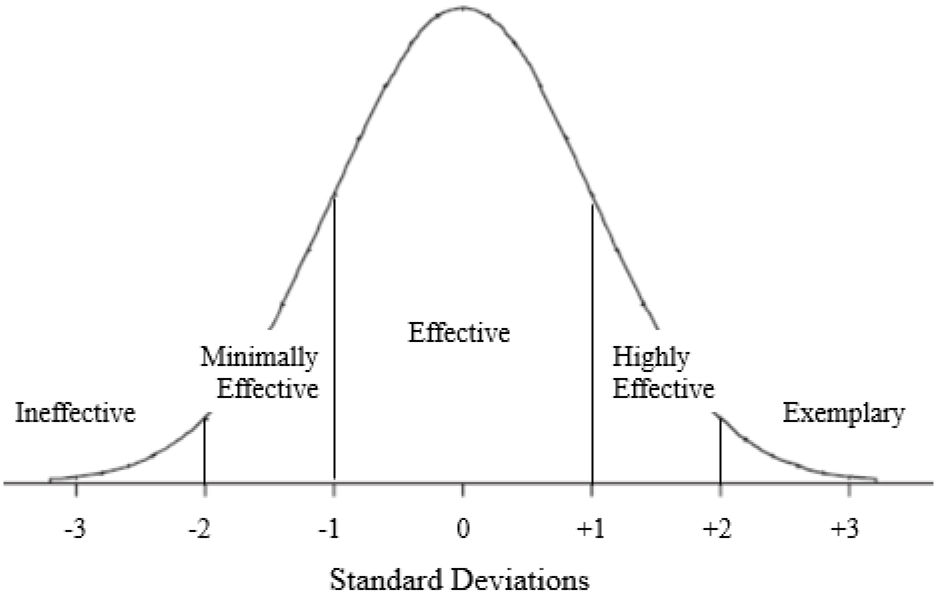


Figure 1. Representation of a normal distribution of teacher effectiveness ratings.

If states succeeded in obtaining more normal distributions of teacher effectiveness, policymakers and others supporting the push for increased teacher accountability would interpret those distributions as indicators that the systems were being properly "reformed" (i.e., as the distributions would better fit the conceptual and perceived distributions of teacher effectiveness [i.e., a bell curve]). Further, and more importantly, states obtaining more normal distributions of teacher effectiveness would also be perceived as being committed to holding teachers accountable, which would subsequently reinforce states' efforts in trying to reform their teacher evaluation systems to achieve more normal distributions. The end result of this increased teacher accountability and the reformed systems would, in theory, result in improved student learning and achievement and, subsequently, the country would regain and maintain its global prominence and superiority.

As a result, teacher evaluation systems as a whole, as well as their individual components, quickly shifted from being mostly formative in nature (e.g., used to inform professional development; Cohen & Goldhaber, 2016; Hibler & Snyder, 2015) to more summative (e.g., used to inform highly consequential personnel decisions, like awarding/denying tenure or merit pay, for example) to better hold teachers (and schools and principals) accountable for their students' achievement. The rationale behind this shift was that if teachers were held accountable, they would be motivated to teach more effectively, and as a result, students would learn and achieve more. Further, if high stakes (e.g., pay, tenure, possible termination) were attached to teachers' effectiveness, teachers would take their roles even more seriously, which would ultimately serve to benefit students' learning and the country as a whole (see Firestone, 2014; Koretz, 2017).

To facilitate the push for accountability, one of the most controversial measures of teacher quality to date, the VAM, fully secured its position within teacher evaluation systems (Collins & Amrein-Beardsley, 2014). VAMs were seen by many policymakers and some groups of researchers as an ideal way to identify effective and ineffective teachers due to a VAM's "objective" (i.e., data-driven) nature (Doran & Izumi, 2004;

Linn, 2004; see also Taylor & Tyler, 2012), as well as *the* measure that would finally help states achieve a more normal distribution of teacher quality.

Value-Added Models (VAMs)

Generally speaking, VAMs aim to statistically measure and then classify teachers' levels of effectiveness based on the impact each teacher has on his or her individual students' achievement over time (Amrein-Beardsley, 2014). VAM modelers typically calculate these teacher effects by measuring student growth over time on standardized tests, and then aggregating this growth at the teacher level, while sometimes statistically controlling for potentially confounding variables such as students' prior test scores and other student-level characteristics (e.g., free and reduced lunch classification [FRL], English language learner [ELL] status, special education [SE] status) and school-level variables (e.g., class size, total school enrollment), although control variables vary by model. Teachers whose students collectively outperform said students' projected levels of growth are identified as teachers of "added value," and teachers whose students fall short of projections are identified as teachers not of "added value."

VAM repercussions. Ultimately, what resulted from widespread VAM use between 2010 and 2015 (i.e., the years immediately following the RTTT initiative), with over 80% of states requiring student achievement data in their teacher evaluation systems (Walsh et al., 2017; see also Steinberg & Donaldson, 2016), was disastrous. Teachers felt angry, frustrated, and hopeless by being evaluated by VAMs (Alm, 2017; Astor, 2018; Collins, 2014), given VAMs' inconsistencies, complexities, and perceived unfairness (described in more detail in Chapter 2, forthcoming), and especially when VAMs were used to inform high-stakes personnel decisions (e.g., teacher termination, awarding/denying merit pay, awarding/denying tenure). In some districts, teachers who were deemed “ineffective” per their VAM scores, were, in essence, publicly shamed in local and national newspapers (*Los Angeles Times*, 2010; see also Gabriel & Lester, 2013b), even driving one teacher to commit suicide (Pathe & Choe, 2013).

In the academy, scholars appeared to be either staunchly in support of or against VAMs (Amrein-Beardsley & Holloway, 2019). VAM detractors cautioned that VAMs were too unreliable, invalid, biased, unfair, or incomprehensible to be used to evaluate teachers, especially for high-stakes purposes (e.g., Amrein-Beardsley, 2014; Baker, Oluwole, & Green, 2013; Hill, Kapitula, & Umland, 2011; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Rothstein, 2009, 2010). Yet VAM supporters felt that the concerns raised about VAMs were based on faulty logic, rendered from poor or incorrect research methods, or, simply put, purely overstated worries (e.g., Adler, 2013; Chetty, Friedman, & Rockoff, 2014a, 2014b; Corcoran & Goldhaber, 2013; Kane & Staiger, 2008; Koedel & Betts, 2007; Papay, 2011; Pivovarova, Broatch, & Amrein-Beardsley, 2014; Rivkin, Hanushek, & Kain, 2005; Sanders & Horn, 1998; see also Tobiason, 2018). Even professional organizations, such as the American Statistical Association ([ASA], 2014; see also Morganstein & Wasserstein, 2014) and the American Educational Research Association (AERA) (AERA Council, 2015) felt the need to chime in on the VAM debate, delivering rare policy statements (e.g., about the proper use of VAMs), and several scholarly journals devoted special issues specifically to VAM use and high-stakes teacher evaluation and accountability policies (e.g., Holloway, Sørensen, & Verger, 2017; Wainer, 2004).

The contentious nature of the summative use of VAMs in teacher evaluation systems ultimately came to a head when teachers and teacher unions began filing lawsuits challenging, in essence, the accuracy and fairness of VAM use in teacher evaluation systems. As of 2015, when VAM use was at an all-time high across the U.S. (Walsh et al., 2017), 14 teacher evaluation lawsuits across seven states had been filed (Sawchuk, 2015). Overall, teacher plaintiffs contested the high-stakes consequences attached to teachers' alleged impacts on their students' test scores over time, including but not limited to being denied merit pay, being denied tenure, being terminated, and other "unfair penalties" for poor evaluation scores.

The Case of New Mexico

New Mexico was one of two states that was celebrated as achieving a "better" (i.e., closer to normal) distribution of teacher quality as per its post-RTTT reformed teacher evaluation system (i.e., beginning in the 2013-2014 school year; Kraft & Gilmour, 2017). As a result of this closer-to-normal distribution, New Mexico was labeled as a "gold standard" state by policymakers and several teacher interest groups (e.g., the National Council on Teacher Quality [NCTQ]). This declaration was solely due to the state's evaluation system's purported ability to objectively and accurately differentiate between its effective and ineffective teachers (Burgess, 2016; Doherty & Jacobs, 2015; Walsh et al., 2017). Christopher Ruszkowski, the New Mexico Secretary of Education from 2017 to early January, 2019, said that the state's system and its ability to differentiate teachers based on effectiveness was a testament to the state's "commitment to putting students first" (Burgess, 2017, para. 6). He also accused other states of "turn[ing] their back[s] on their commitments to [students]…[by] painting a picture [of the distribution of teacher quality] that we know is not accurate" (i.e., a negatively skewed distribution) (paras. 10-11).

Compared to other states, New Mexico's teacher evaluation policy put a bigger emphasis on students' test scores (i.e., via VAMs) (Doherty & Jacobs, 2015), likely in large part to counter the negatively skewed distribution of teachers' classroom observation scores and student survey scores. The combination of measures, and, specifically, the weighting of each measure resulted in a nearly normal distribution of teachers' overall effectiveness ratings (see Figure 2 for a representation of what this might look like; see also Figure 3 in Chapter 4).

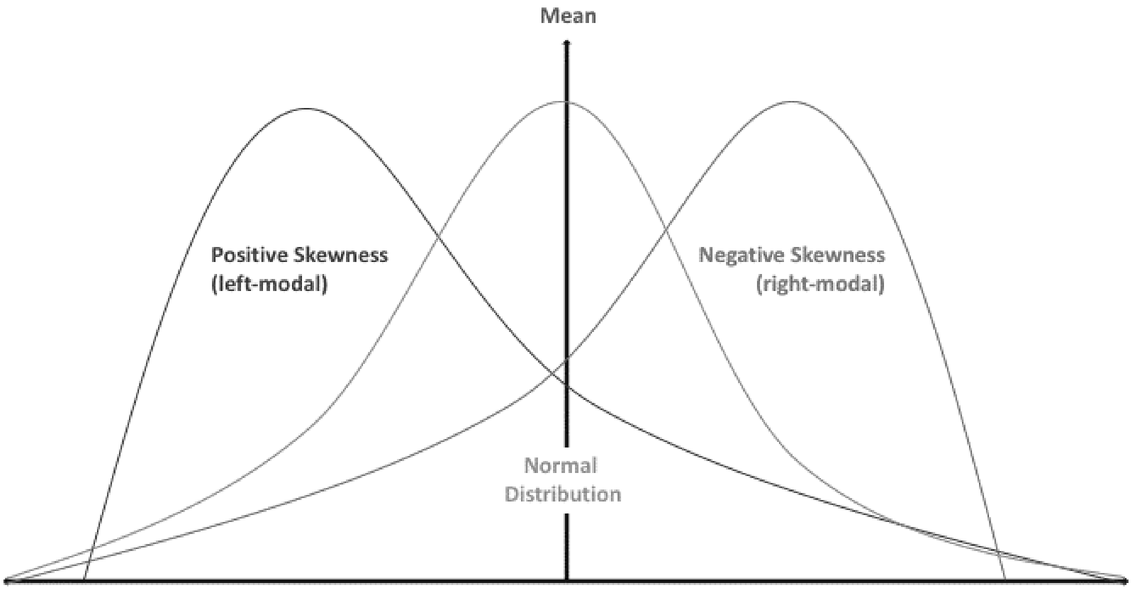


Figure 2. Depiction of how weighting skewed data can yield a normal distribution

While this distribution was celebrated by New Mexico policymakers and others (e.g., Doherty & Jacobs, 2015), many teachers throughout the state publicly condemned the system, citing its heightened focus on students' test scores (Frosch, 2013; Heinz, 2011; Nadeem, 2013). Teachers also reported feeling "devalued" (Burgess, 2017, para. 21) by this system and saw it as "extreme" and "out-of-touch" (para. 22). Ultimately, the perceived unfairness of the state's system by teachers, along with claims that too many teachers were being classified as ineffective, largely due to VAMs and the engineering of a normal distribution of effectiveness ratings, led to three of the country's 14 teacher evaluation lawsuits.

In early 2015, a group of New Mexico educators, stakeholders, and politicians[1] filed what was the third teacher evaluation related lawsuit in the state. The plaintiffs sued the New Mexico Public Education Department (NMPED) and the state's then-education secretary designate, Hanna Skandera (*State ex rel. Stewart v. New Mexico Public Education Department*, 2015), alleging that New Mexico's teacher evaluation system was detrimental to both teachers and students and that it violated the state's requirement for all teachers to be evaluated by "highly objective uniform standards" (New Mexico Administrative Code [NMAC], 2011; see also American Federation of Teachers, 2015). Specifically, the plaintiffs claimed that 1) New Mexico teachers received poor value-added scores (VAS; i.e., VAM scores in New Mexico) due to flawed and incomplete student-level data (e.g., teachers were linked to the wrong students, students they never

[1] The plaintiffs consisted of the American Federation of Teachers – New Mexico (AFTNM), the Albuquerque Teachers Federation (ATF), six Albuquerque teachers, one Gallup teacher, four New Mexico senators, and one New Mexico representative.

taught, subject areas they never taught, or using tests that did not map onto that which they taught); and 2) the consequential decisions (e.g., flagging teachers' files, teacher termination decisions) that were informed mostly by teachers' VAS scores were arbitrary and not legally defensible, in that they were statistically unreliable, invalid, and biased.

After this third lawsuit was filed, District Judge David K. Thomson granted a preliminary injunction that prevented New Mexico from making any further high-stakes decisions about its teachers using the then-current teacher evaluation framework, unless the state could provide evidence to the court that the teacher evaluation system as a whole and its individual measures were all reliable, valid, and unbiased (see Amrein-Beardsley, 2018a). As a result, in early 2018, Dr. Audrey Amrein-Beardsley was called upon by the plaintiffs' lawyers to act as the lone expert witness for the plaintiffs, given her expertise in VAMs, teacher evaluation systems, and related accountability policy. She was tasked with analyzing the state's teacher evaluation system to determine if it was indeed "highly objective and uniform," as state policy required (NMAC, 2011). The specific analysis goal was to analyze data for each of the four measures of teacher effectiveness (i.e., VAS scores; classroom observation scores; Planning, Preparation, and Professionalism [PPP] scores; and SPS scores) from the 2013-2014, 2014-2015, and 2015-2016 school years to determine each measure's levels of statistical reliability, validity, and bias (or lack thereof). There was to be a concerted focus on teachers' VAS scores, as they were the most heavily weighted component of the teacher evaluation system and also given their notable contentious attributes (described in more detail in Chapter 2).

Analyses for the *State ex rel. Stewart v. New Mexico Public Education Department* (2015) lawsuit. As Dr. Amrein-Beardsley's Research Assistant at the time of her role as expert witness, I performed the majority of the requested analyses by examining whether and to what extent each of the aforementioned teacher evaluation measures showed indications of (un)reliability, (in)validity, and bias (or lack thereof). Results from those analyses indicated that, overall, teachers' VAS scores, along with the other measures of teacher effectiveness, were not consistently reliable, valid, or unbiased and therefore violated the state's "highly objective and uniform" clause (Amrein-Beardsley & Geiger, revise and resubmit; see also Amrein-Beardsley, 2018b).

When I performed those analyses, I did so under a narrow scope based on the specifics of the request from the plaintiffs' lawyers. Specific to the question of bias, while I performed several inferential statistical tests (e.g., *t*-tests, ANOVA) that led to the aforementioned findings, I did not control for any extraneous factors (i.e., covariates; e.g., student or teacher demographic factors) that might have affected both the results of the statistical tests and the conclusions that Dr. Amrein-Beardsley and I drew from those results. As such, I felt that subsequent analyses into the New Mexico teacher evaluation system measures were warranted to determine if the previous findings help up under more robust methods.

Study Purpose

Although the District Judge in the *State ex rel. Stewart v. New Mexico Public Education Department* (2015) granted a preliminary injunction that forbade the NMPED from making any further consequential decisions using its then-current teacher evaluation framework, New Mexico still continued to use VAMs to evaluate teachers, also notwithstanding the plethora of researchers and scholars documenting the many concerns about VAMs (discussed in more detail in Chapter 2, forthcoming). In theory, as previously mentioned, VAMs are supposed to control for potential biasing factors, including but not limited to student background characteristics such as a student's SE, ELL, FRL, and underrepresented minority (URM) status (among others). However, prior researchers (e.g., Ballou, Sanders, & Wright, 2004; Ehlert, Koedel, Parsons, & Podgursky, 2016; Fuller, 2014; Kane, 2017; Newton et al., 2010; Michelmore & Dynarski, 2017) have suggested that such background characteristics have been significantly associated with teachers' VAM scores.

These associations are especially problematic because students are never randomly assigned to schools (or classrooms) (see, for example, Lomax & Hahs-Vaugh, 2012; Paufler & Amrein-Beardsley, 2014; Rothstein, 2009, 2010). That is, overall, certain types of students (e.g., SE, FRL, ELL, and/or URM students) tend to be clustered in certain schools more so than others. This lack of random assignment will likely never change given, for example, the purposeful enrollment (or avoidance) of certain students in certain schools (i.e., in states with open enrollment, like New Mexico [New Mexico Statute §22-1-4]) or residential segregation (e.g., Frankenberg, 2013; Quillian, 2014).

Thus, the purpose of this study was to examine whether and to what extent student background characteristics, aggregated to the school level, affected New Mexico's teacher evaluation measures during the 2013-2014, 2014-2015, and 2015-2016 school years. Specifically, I investigated whether and to what extent the student composition within schools, based on four notable student background characteristics (i.e., SE status, ELL status, FRL status, and URM status), affected teachers’ VAS, observation, PPP, and SPS scores in New Mexico. In general, if VAMs or other measures of teacher effectiveness are significantly associated with student background characteristics, teachers who work in certain schools risk being evaluated unfairly. These unfair evaluations might look like teachers being identified as less effective than they truly are, and therefore possibly penalized. The converse is also possible: teachers might be identified as more effective than they truly are, and therefore possibly not receive much needed professional development, or be disciplined or terminated. Both of these scenarios are detrimental to the country’s students, the teachers themselves, and the American education system writ large.

Research Questions

Drawing from same New Mexico’s teacher effectiveness dataset used in the *State ex rel. Stewart v. New Mexico Public Education Department* (2015) lawsuit, I answered the following overarching research questions: 1) What are the relationships between student background characteristics, aggregated to the school level, and the four main teacher evaluation measures that comprised a teacher’s overall evaluation score in New Mexico during the 2013-2014, 2014-2015, and 2015-2016 school years? and 2) How do these relationships compare across the four main teacher evaluation measures?

To answer these two main research questions, I answered the following 16 sub-questions (which are grouped by teacher evaluation measure for clarity):

1. Value-Added (VAS) scores:

1a.) What is the relationship between the percent of SE students within a teacher’s school and the percent of VAS points a teacher earns?

1b.) What is the relationship between the percent of ELL students within a teacher’s school and the percent of VAS points a teacher earns?

1c.) What is the relationship between the percent of FRL students within a teacher’s school and the percent of VAS points a teacher earns?

1d.) What is the relationship between the percent of URM students within a teacher’s school and the percent of VAS points a teacher earns?

2. Classroom observation scores:

2a.) What is the relationship between the percent of SE students within a teacher’s school and the percent of observation points a teacher earns?

2b.) What is the relationship between the percent of ELL students within a teacher’s school and the percent of observation points a teacher earns?

2c.) What is the relationship between the percent of FRL students within a teacher’s school and the percent of observation points a teacher earns?

2d.) What is the relationship between the percent of URM students within a teacher’s school and the percent of observation points a teacher earns?

3. Planning, Preparation, and Professionalism (PPP) scores:

3a.) What is the relationship between the percent of SE students within a teacher’s school and the percent of PPP points a teacher earns?

3b.) What is the relationship between the percent of ELL students within a teacher’s school and the percent of PPP points a teacher earns?

3c.) What is the relationship between the percent of FRL students within a teacher’s school and the percent of PPP points a teacher earns?

3d.) What is the relationship between the percent of URM students within a teacher’s school and the percent of PPP points a teacher earns?

4. Student Perception Survey (SPS) scores:

4a.) What is the relationship between the percent of SE students within a teacher’s school and the percent of SPS points a teacher earns?

4b.) What is the relationship between the percent of ELL students within a teacher’s school and the percent of SPS points a teacher earns?

4c.) What is the relationship between the percent of FRL students within a teacher’s school and the percent of SPS points a teacher earns?

4d.) What is the relationship between the percent of URM students within a teacher’s school and the percent of SPS points a teacher earns?

Study Significance

While decades of research have been conducted on teacher effectiveness and teacher evaluation measures, including research specifically on VAMs, from a multitude of perspectives (e.g., measurement, policy, (un)intended consequences), this study contributes to the growing body of literature about the ways in which teachers might best be evaluated (or not). Specifically, via this study, I provide an in-depth look into the measures that were used between 2013-2016 to evaluate teachers in New Mexico. To date, I believe that only one other published study used contemporary New Mexico teacher evaluation data (see Doan, Schweig, & Mihaly, 2019). While New Mexico has since abolished VAMs (State of New Mexico, 2019), multiple other states across the country still continue to use them to evaluate teachers (Close, Amrein-Beardsley, & Collins, 2018), even though they are no longer federally mandated (Every Student Succeeds Act [ESSA], 2015).

This study is also of note as it is rare for a researcher to have access to an entire state's teacher evaluation dataset, including with corresponding (albeit limited) teacher and student demographic data, and across multiple years. Thus, this study represents one of the few instances where such a large dataset has been utilized to evaluate teacher evaluation measures. Lastly, this study takes further conceptual significance when one considers how New Mexico has been seen, for years, as a "gold standard" state for having one of the "best" teacher evaluation systems in the country—that is, a system that produces a nearly normal distribution of teacher effectiveness ratings (Doherty & Jacobs, 2015; Kraft & Gilmour, 2017; Putnam, Ross, & Walsh, 2018). The apparent contradiction between the state's "gold standard" teacher evaluation system and multiple lawsuits about the unfairness of that very system make this context and therefore this study especially unique.

Until federal policy expressly outlaws the use of VAMs in any and all high-stakes situations and/or for summative purposes, or until VAMs are drastically overhauled from a methodological and measurement standpoint (and to what extent that is possible is likely unknown), VAM consumers—namely federal and state policymakers; state, district, and school administrators; and teachers—need to better understand the measurement and pragmatic issues at hand, especially in high-stakes contexts. Findings from this study will be important not only for others who are still grappling with whether and how to use VAMs to evaluate and hold teachers accountable for that which they do or do not do well, but also for states, districts, and schools that continue to adopt, implement, or even merely consider VAMs in their teacher evaluation systems.

Overview of the Dissertation

In Chapter 2, I provide an in-depth overview of the history of educational accountability policy with a focus on teacher evaluation, both federally and specific to the state of New Mexico. I then discuss the history of each of the three main measures used to evaluate teachers in New Mexico during the 2013-2014 through 2015-2016 school years (i.e., VAMs, classroom observations, SPSs), and also include a summary of the current literature about teach measure. Lastly, I close with an explanation of the conceptual framework that I used to situate this study.

In Chapter 3, I explain the methodology used to answer this study's research questions. I provide details about my data source and samples of participants, my data cleaning process, and explain the rationale for using multiple linear regression. I end this chapter with a discussion of the study's limitations.

In Chapter 4, I present the study results. I begin with describing my samples, overall and per year, and then present the results from each of the 48 regression models. I also provide brief summaries of each set of models (i.e., as grouped by teacher evaluation measure). I close this chapter with a brief summary of the overall results.

In Chapter 5, I discuss the findings that stemmed from the study’s results. I offer specific insights into these findings as relevant to each of the four measures and overall. I also situate these findings within the current literature. I end this chapter by presenting two possible interpretations of this study’s findings.

In Chapter 6, the final chapter, I first briefly summarize the study. I then present and discuss three possible implications stemming from this study’s findings as related to both an applied perspective (i.e., the measures used to evaluate teachers) and a theoretical one (i.e., the rationale used to push for a normal distribution of teaching quality). Lastly, I close the dissertation by addressing directions for future research and inquiry.

CHAPTER 2

LITERATURE REVIEW

In this chapter, I explain the history of federal education policy as related to accountability, with a focus on teacher accountability and teacher evaluation policies. I then outline the history of New Mexico's education accountability policies, also with a specific focus on teacher accountability and teacher evaluation policies. I then discuss the history of and empirical research about the three teacher evaluation measures used in New Mexico between the 2013-2014 and 2015-2016 school years: VAMs, classroom observations, and student perception surveys (SPSs). Lastly, I describe the conceptual framework that undergirds this study.

History of Teacher Evaluation in the United States

Elementary and Secondary Education Act and the Coleman Report

In January 1965, the Elementary and Secondary Education Act (ESEA) was passed by the Lyndon Johnson administration to improve the quality of education and educational opportunities in the U.S. Prior to ESEA, the federal government had little oversight when it came to education, as decisions were often made at the state or local level. ESEA was enacted under the premise that it would provide funding to school districts to improve outcomes and opportunities for disadvantaged students. Specifically, Title I of ESEA called for funding to be provided to schools that had high proportions of low-income students. ESEA was significant because it was the first legislation of its kind where the federal government was involved in aiding education (Thomas, 1983), and it also began the modern accountability movement by cementing the federal government's interest in addressing disparities and inequities in outcomes and opportunities for disadvantaged students (Kantor, 1991). Shortly after ESEA was passed, the Coleman Report (Coleman et al., 1966) was released, which detailed the first large-scale study that focused on the potential effects that a variety of inputs (i.e., school- and teacher-level factors) had on student achievement. The Coleman Report has been recognized as one of the most pivotal studies conducted in the 20th century as it debunked the idea that school quality was the most influential factor in students' academic achievement and instead reported that students' family characteristics, such as socioeconomic statuses, were more accurate predictors (Hoff, 1999). The Coleman Report (Coleman et al., 1966), in combination with ESEA's focus on improving education for disadvantaged students, paved the way for future social science research on educational outcomes and opportunities for students (Wong & Nicotera, 2004).

In the late 1960s and 1970s, there were few policies, if any, that focused on accountability as we conceptualize it today. Teacher evaluation policies were often left to be decided by individual schools or districts (rather than at a state or federal level), and there were no real policies that focused on student-level or school-level accountability. During this time, the majority of teacher evaluations were based on classroom observations of teachers, though educational and other social science researchers were trying to determine what relationship, if any, existed between effective teaching and student outcomes (Ellett & Teddlie, 2003). While the Coleman Report (Coleman et al., 1966) findings were more noteworthy regarding non-school factors affecting student achievement, Coleman and colleagues also noted that teacher quality was the most impactful school-level factor on student achievement. In the 1970s, more studies (e.g., Hanushek, 1970, 1979) were conducted to look into further detail about exactly how schools and teachers affected student achievement.

Minimum Competency Era

During the early to mid-1970s, policymakers and the general public had become more displeased with the state of U.S. public education as achievement gaps between students based on race and socioeconomic status were prevalent, national test scores had been declining, and both the achievement gaps and the test score decline became highly publicized (Haertel & Herman, 2005). Policymakers had become less focused on improving resources and curriculum and more focused on paying closer attention to student outcomes (Haertel & Herman, 2005), especially as reports of thousands of ill-prepared students graduating from high school surfaced (Cole, 1979).

Further, in the 1970s, policymakers realized that current standardized tests could be utilized as a method of accountability in that students' scores could be used to quantify how much students were learning (Behuniak, 2003). This was quite different than the previous uses of standardized tests, which were typically used in a formative sense for teachers (Behuniak, 2003). The aforementioned achievement gap and test score decline; the draw of accountability; and the tenets of behaviorism and social psychology, which had begun to gain traction in the one or two decades prior (Ellett & Teddlie, 2003), helped guide the nation down the path of minimum competency testing (MCT) and competency-based teacher education (CBTE). CBTE also grew out of the belief that teacher education programs, as then-currently structured, were not generating enough teachers who could properly instruct high-needs students (Elam, 1971). While MCT initially began as a method to assess students' basic skills, it grew into being utilized by schools and districts to ensure teachers were knowledgeable enough to be employed. The following paragraphs mostly focus on MCT (and CBTE) related to teachers and teacher effectiveness.

MCT was initially a method to determine whether students had reached a minimal level of knowledge, typically in reading and writing (Haertel & Herman, 2005). In the late 1970s, states began to require students take such MCTs to ensure they had learned enough to make them, essentially, "competent" in basic skills, or that they had met a minimum standard of knowledge and performance (Houston, 1974). To determine this competency, students' test scores were compared to an established standard (i.e., the minimum competency; Behuniak, 2003). MCT had been put into place in 29 states by 1980 and in 33 states by 1985, and 11 of those 33 states required students to pass a MCT test to receive their high school diploma (Haertel & Herman, 2005).

While MCT was enacted as a student-level accountability measure, the education community and, specifically, state and local school boards, were also looking to ensure that newly hired teachers had enough basic knowledge and general teaching skills to be effective in the classroom (Harris, 1981). Harris explained that the logic behind this supplementary argument was students being deficient in certain subject matters or skills was also a result of their teachers being similarly inadequate. Thus, CBTE (or, similarly, performance-based teacher education [PBTE]) was born. CBTE was supported by the following assumptions: teachers who could demonstrate minimum competencies were more effective than those who could not; teacher quality would improve if programs like CBTE (i.e., where teachers must demonstrate competence) were implemented; and after such implementation, the reputation of the teaching profession would improve if it was known whether teachers passed a test to be certified to teach. To do this, a cut score or “critical point” for competencies would be identified, also given teacher preparation programs had previously done a poor job at screening out unqualified candidates for the profession (Pugach & Raths, 1983).

While the promise of CBTE was high, many in the education sector had immediate concerns, especially as CBTE quickly became the norm for many teacher preparation programs. Notably, issues included how to agree upon and define effective teaching behaviors or competencies (Benham, 1981), how to reliably measure such competencies (Quirk, 1974), how to properly simulate teaching in schools in a preparation program (Elam, 1971), and how to handle the lack of necessary funds needed to institute and maintain such programs (Elam, 1971; Jarrett, 1977). Although CBTE was initially highly popular and the next big “coming thing” (Houston & Howsam, 1972, p. 4), for accountability purposes it quickly lost its steam in the early 1980s. In a summary of CBTE programs, Roth (1977) concluded that although some programs showed promise, overall, it was not possible to determine to what extent, if at all, CBTE programs were effective. Teachers and other stakeholders felt that teaching as a profession was more complicated than merely demonstrating proficiency on specified competencies (Broudy, 1972; Darling-Hammond & Wise, 1985); further, Piper and

Houston (1980) noted that the actual outcome of CBTE was improving teacher performance rather than holding teachers accountable for their performance.

A Nation At Risk

As the focus on basic skills and competencies waned in the late 1970s and early 1980s, educational "reform" and evaluation through test-based accountability gained traction. In 1981, then-Secretary of Education T. H. Bell created the National Commission on Excellence in Education (NCEE) to assess and respond to concerns, via an official report, about growing worries about the quality of American education. The NCEE (1983) investigated six main areas pertaining to the American educational system: (1) the quality of teaching and learning occurring in all public and private schools at all levels, (2) how U.S. schools compared to schools of other industrialized nations, (3) the relationship between students' high school academic achievement and college admission requirements, (4) programs that result in "student success" in college, (5) the relationship between social and educational changes in the past 25 years (i.e., since ESEA) and student achievement, and (6) problems that need to be solved to pursue educational excellence as a nation.

The subsequent report, titled "A Nation at Risk," painted an ominous, alarming, and dire picture of the state of American education—with the title being a clear indication of the current status of education in the country. The report's main message was that the U.S. was being threatened by a "rising tide of mediocrity" (NCEE, 1983, p. 12), the educational system in the country was failing, and thus, without improvement, the U.S. would cease to be competitive on the global stage against other industrialized nations.

This idea was a deviation from previous messages from the federal government, as *A Nation at Risk* focused on improving the educational system in its entirety, rather than just focusing on certain subgroups of students (e.g., disadvantaged students, those affected by the achievement gap, and the like) (Birman, 2013).

The report focused on four main elements of education and schooling—academic content, use of time, expectations, and teaching—and provided four main recommendations. For academic content, the NCEE (1983) recommended that to earn a high school diploma, students take four English courses, three mathematics, science, and social studies courses each, and a half credit of computer science. Students planning to pursue a college education were also recommend to take two credits of a foreign language. Regarding time, the NCEE recommended that schools allot more time to teaching basic subjects and curriculum, which would result in either longer school days, a longer school year, or a more efficient use of time in the current schedule.

The recommendations for improved expectations and teacher quality were a large focus of the discourse and rhetoric across the national education landscape after the report was released. The NCEE recommended that all schools have higher expectations for student achievement along with rigorous and measurable standards to assess performance. The recommendations for improving teacher quality were extensive, as the report pinpointed teachers as the main group who had the ability and potential to shape the U.S.'s future workforce (as the country's economy, national security, and international competitiveness were at risk due to the "failing" of American education). Specifically, the NCEE recommended more rigorous standards for teacher preparation programs; improved teacher pay, professional development, mentoring programs, and recruitment incentives; longer contracts; and improved career ladders.

Overall, the NCEE's recommendations called for the strengthening of the federal government's role in improving schools (McGuinn, 2006). While the report was frequently cited by then-President Reagan and its ideas gained steam among the American public, many education scholars were quite critical. Some critics saw the allegation that the U.S. had a mediocre or failing educational system as a "manufactured crisis" (Berliner & Biddle, 1995), or a scare tactic used to try to improve American education. These critics believed that the NCEE's (1983) claims about declining student performance, poor school quality, and a weakening American workforce were incorrect (see also Guthrie & Springer, 2004). While there were plenty of criticisms of the report and commentaries about its flaws (see also Goodlad, 2003; Peterson, 2003), it was still widely circulated and continued to gain enough attention that the media, policymakers, and the general public believed its message. Further, other complementary reports followed (e.g., *Action for Excellence: A Comprehensive Plan to Improve our Nation's Schools* [Education Commission of the States, 1983]), all with similar messages about the state of American education.

At the same time, during the mid to late 1980s, general discourse about American education was taking a negative tone in that the country's educational system needed to be revamped. Many believed that the federal government's oversight via ESEA and, specifically, Title I of ESEA, was too strict, the various program requirements were too confusing and complex to meet the goals of reducing the achievement gaps among certain subgroups of students, and thus, the law was actually a hindrance to educational progress for the very students it was intended to help (Birman, 2013). *A Nation at Risk* and the growing discontent of ESEA, Title I, and its subsequent programs led to the foundation of the standards-based reform movement that commenced in the 1990s (Weiss, 2003), and many states created policies that used testing as a measure of student achievement and as a means to evaluate teachers (Koretz, 1992, 1996). One assessment that lent itself well to the beginning of the standards-based movement was the National Assessment of Educational Progress (NAEP), a test previously created to measure student academic performance in a variety of subject areas. Although the NAEP was created during the mid to late 1960s, it was first used at the statewide level on a voluntary basis in 1990, after which participating states received their own data as well as other states' results for comparison purposes (Beaton et al., 2011).

At the same time as the NAEP was being piloted in states on a voluntary basis, and in the few years that followed, the federal government (again) began to discuss how to improve education through promoting high standards for all students; improved teacher preparation programs and teacher training; providing flexibility to serve as a catalyst for local reform with accompanying accountability measures; and creating partnerships among families, communities, and schools (Riley, 1995). These focal points were prevalent in the next wave of education legislation: the 1994 reauthorization of ESEA (i.e., the Improving America's Schools Act) and the Goals 2000: Education America Act.

Improving America’s Schools and Goals 2000

The 1994 reauthorization of ESEA, known as the Improving America’s Schools Act (IASA), came with the shift to the fervent beginning of the rhetoric and actions of standards-based reforms and accountability. IASA focused on high standards for all students; professional development for teachers and administrators to support students reaching high standards; flexibility for states, districts, and schools to implement federal programs as they saw fit, along with accountability standards to measure implementation and results; and promoting partnerships between schools and communities (Riley, 1995). The initial ESEA was criticized for being too stringent, inflexible, and difficult to adequately measure, so IASA and specifically, Title I, contained several changes that attempted to remedy those issues.

First, Title I of the ESEA reauthorization directed states to set high curriculum and student performance standards to receive federal funding (Birman, 2013), and specified that future funding was dependent on this development (Stedman, 1994b). This connected previously-created programs under ESEA to systemic reforms and increased accountability and transparency into these programs. States, districts, and schools were also given increased flexibility by transitioning previous ESEA programs from the federal level to the state level, which allowed states to implement programs in ways that best fit their respective needs. The USDOE was also allowed to waive a plethora of ESEA requirements for a state if that state could provide evidence that such a waiver would help increase student performance. The IASA also created several programs and initiatives that focused on new emerging areas of interest, such as technology, school safety, and the alternative management of schools. Lastly, Title I funding allocation formulas were adjusted to ensure that students in schools that were the most economically disadvantaged received needed funds (Stedman, 1994b).

Related to IASA was the Goals 2000: Educate America Act of 1994, legislation that was known for being at the center of then-President Clinton's education platform. Goals 2000 was a list of eight national objectives set by Congress that were based on the ideas of standards-based reform. While the act created national standards, the onus was on states to develop and implement the necessary reforms to meet the stated objectives (Schwartz & Robinson, 2000). The intent of the act was for all of the following goals to be met by the year 2000:

- All children will start school ready to learn;
- The high school graduation rate will be at least 90%;
- All students will leave grades 4, 8, and 12 having demonstrated competency over challenging subject matter…and every school in America will ensure that all students learn to use their minds well…;
- [Teachers] will have access to programs for the continued improvement of their professional skills and the opportunity to acquire the knowledge and skills needed to instruct and prepare all American students for the next century;
- U.S. students will be first in the world in mathematics and science achievement;
- Every adult American will be literate and will possess the knowledge and skills necessary to compete in a global economy and exercise the rights and responsibilities of citizenship;

- Every school…will be free of drugs, violence, and the unauthorized presence of firearms and alcohol…; and
- Every school will promote partnerships that will increase parental involvement and participation in promoting the social, emotional, and academic growth of children. (Stedman, 1994a, pp. 7-8)

In addition to the eight goals, Goals 2000 also created several groups (e.g., the National Education Goals Panel, the National Education Standards and Improvement Council, the National Skill Standards Board) to monitor goal progress and certify assessments; established and certified a variety of national and state standards, such as those related to curriculum, performance, and assessment; provided the ability to waive certain requirements created under previous federal education programs; and appropriated the necessary funds (of over $105 million) to states that developed and implemented plans and reforms that supported reaching the aforementioned goals (Stedman, 1994a). While participation in Goals 2000 was voluntary for states, opting out of developing and implementing reforms related to the eight goals resulting in states waiving their right to this federal funding.

Goals 2000 was the first legislation that did not focus on providing an equity-based solution for underserved and disadvantaged students and schools (as had all previous legislation) but rather focused on improving the academic performance of all students at all schools (McGuinn, 2006), as well as officially shifting education-related policies from local authorities to the federal government (McGuinn, 2006; Schwartz & Robinson, 2000). On the surface, it seemed that Goals 2000 helped states—and therefore the nation—make progress in terms of providing and improving high quality education for all students. However, by 2000, many states and districts had yet to even begin to implement the far-reaching changes supported by the legislation, and many states faced challenges with implementing their proposed reform measures and programs (Schwartz & Robinson, 2000). Goals 2000 ultimately fizzled out as there was no way to ensure compliance with the reforms outlined in the act. However, from a policy regime framework that was now hyper-focused on standards-based assessment and accountability, Goals 2000 paved the way for NCLB—one of the most impactful pieces of education policy in the past 20 years. Additionally, although teacher evaluation was not at the forefront of education policy during the late 20th century, the focus on student performance was setting the stage for evaluating teachers via students' academic achievement.

No Child Left Behind (NCLB)

NCLB was the 2001 reauthorization of ESEA and its direct focus was on improving student performance outcomes via a variety of accountability measures and mandates, with strict consequences for states (and schools and districts) that missed hitting outlined goals. NCLB's main purpose was for all students to have "a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging state academic achievement standards and state academic assessments" (p. 15). Per the USDOE (2003), NCLB was created on four "common-sense pillars" (p. x): accountability, as related to student academic achievement and teacher quality; expanded local flexibility and control, in terms of developing and implementing standards, assessments, and how to best spend federal funds; increased parental options; and utilizing effective programs that have been substantiated via scientifically based research.

NCLB outlined several areas for its sweeping reforms, including the development and implementation of high quality assessments to appraise student academic performance with corresponding data and accountability systems; improved teacher preparation and training programs and the mandate that all teachers in all schools must be deemed “highly qualified;” and improved local flexibility and control for schools and districts, such as determining how to best use federal funding (Simpson, LaCava, & Graner, 2004). Each state and district was required to produce an annual report card that listed whether each school was succeeding. This report card also needed to include student achievement data at both the aggregate and disaggregated levels, as well as teachers’ qualifications (USDOE, 2004). The following subsections further detail the accountability mandates as related to student performance and teacher quality, both of which set the stage for later efforts to hold teachers accountable for their students’ performance.

Student performance. NCLB stipulated that all states had to create and utilize standardized tests along with accompanying data systems to assess students’ academic performance, and also disaggregate and analyze students’ test scores by a variety of subgroups (i.e., race/ethnicity, poverty, disability status, etc.) to ensure that, literally, no students were “left behind” in terms of their academic achievement. The federal government believed that such testing plans would provide insight into each student’s, school's, district's, and state's progress and achievement, and would therefore ensure that students were not "trapped" (USDOE, 2003, p. 11) in poor-performing schools. Stated additional benefits of this increase in testing were the identification of "problem areas," such as aspects of school curricula that needed to be reviewed and/or better aligned with state content standards or instructional methods for which teachers could improve or use professional development (USDOE, 2003).

Per the NCLB testing plan, every state needed to assess every student in both reading and mathematics between grades three through eight, and at least once in grades 10 through 12 by the 2005-2006 school year. Further, by the 2007-2008 school year, each state would also need to assess student performance in science once during grades three through five, once during grades six through nine, and once during grades 10 through 12. In addition to these state-level tests, students in the fourth and eighth grades were also required to participate in the NAEP reading and mathematics assessments every two years, beginning in the 2002-2003 school year. A state that did not include plans to participate in, at a minimum, the required NAEP testing in reading and mathematics in grades four and eight would forfeit their Title I funding (USDOE, 2005a), which most, if not all, states could not allow to happen. This additional NAEP testing also allowed for state-by-state comparisons of student achievement, as well as for rigor and/or content comparisons between students' achievement on the state-level and NAEP tests (USDOE, 2003).

Adequate Yearly Progress. Schools', districts', and states' progress in terms of achievement—and the determination if NCLB's objectives were being met— was to be ascertained by a new metric known as Adequate Yearly Progress (AYP). Per NCLB, each state was supposed to reach 100% AYP by the 2013-2014 school year—meaning each and every student would show, literally, adequate yearly progress, in his or her academic achievement. As a part of the flexibility of NCLB, each state was allowed to determine its own equation for AYP, though school-level AYP was supposed to be a function of academic assessments, graduation rates (if applicable based on the school's grades served; e.g., high school), and other metrics, such as grade-to-grade retention rates or attendance rates, if needed (USDOE, 2005c). While seemingly beneficial on the surface, this so-called flexibility also meant that some states could set their AYP bars too high (or too low) and they would, therefore, have more (or less) work to do to hit the 100% threshold than other states (Rentner et al., 2003). Failure of a school reaching AYP, especially for multiple consecutive years, would result in swift and potentially severe and dire consequences for not just the schools, but also sometimes for the teachers and staff members employed at those schools.

Schools and districts were to be deemed as "satisfactory" if they met AYP in a given year, though this label was to have more of a signaling effect to parents and the general public that the school was deemed to be of high quality. Schools and districts that successfully achieved their AYP goals frequently received public recognition, and teachers and other staff members became eligible to receive public acknowledgement along with fiscal rewards (Simpson et al., 2004). If a school failed to meet AYP for two consecutive years, it would be labeled as "in needs of improvement" and thus had to develop a plan to "turn around," or improve itself. In addition, more important was that students who were enrolled at an “in needs of improvement” school had to be given the option to transfer to another public school in the district. If a school failed to meet AYP for three consecutive years, students still had to be given the transfer option and, further, students from low-income families were allowed to receive additional services, such as tutoring or remedial classes, to help bolster their achievement. If a school failed to meet AYP for four consecutive years, in addition to school choice and supplemental services, corrective action had to be implemented (e.g., replacing staff, implementing a new curriculum). Most dire, if a school failed to meet AYP for five consecutive years, the school had to be completely restructured—meaning it could be reopened as a charter school, the majority of or all the teachers and administrators had to be replaced, or the day-to-day school operations had to be handed over to either the state or a private company that had a record of school operations/management success.

Failing to meet AYP was not just publicly embarrassing, but also costly as a school designated as “in needs of improvement” was required to spend at least 10% of its Title I funds on teacher improvement (e.g., professional development), which was an increase from the typical 5% allocation for successful schools (USDOE, 2004). However, while possibly intimidating on paper, the aforementioned consequences were often inconsistently applied as holding schools accountable for the numerous rules and regulations of NCLB was difficult for the government due to a lack of resources (McGuinn, 2011). Therefore, a school being labeled as “in needs of improvement” was simply more of a negative signaling effect than anything else.

Highly qualified teachers. Another aspect of the NCLB accountability focus was that every classroom in every school had to be staffed with what NCLB labeled a "highly qualified teacher" (HQT). The HQT mandate stemmed from research that positively linked student achievement to teacher quality (Darling-Hammond, 2000; Darling-Hammond & Youngs, 2002; Ehrenburg & Brewer, 1994; Wayne & Youngs, 2003). At a minimum, teachers were "highly qualified" if they had bachelor's degrees, were fully licensed or certified by the state in which they were teaching, and could prove knowledge of the subject they were teaching (U.S. Department of Education, 2005b). To receive Title II funds, which supported preparing, training, and recruiting high quality teachers, states had to create a plan that outlined how they would ensure the HQT requirement would be met. States also had to report what additional indicators they would use to assess whether their teachers were highly qualified, and if their schools or districts were adequately staffed with such HQTs. All teachers in all public schools who taught core subjects (e.g., reading, mathematics, science), as well as those who taught ELL students or students with disabilities were expected to be "highly qualified" by the 2005-2006 school year. Lastly, since previous research had demonstrated a positive relationship between teacher quality and student achievement, and since the USDOE believed so strongly in the HQT provision, NCLB also supported teachers gaining entry into classrooms through alternative certification programs rather than traditional teacher training programs.

The HQT mandate was one of the more controversial aspects of NCLB, as it was costly, difficult to implement, and not fully supported by research (see discussion, forthcoming). However, several years after NCLB was first enacted, due to frequent and consistent criticisms and complaints, the USDOE amended the initial HQT requirements to allow rural schools as well as science teachers and teachers teaching multiple subjects more time to meet the HQT requirement (Simpson et al., 2004).

NCLB reactions. Initial thoughts about and reactions to NCLB were quite varied. Some thought NCLB would dramatically improve the American education system as a whole (Simpson et al., 2004), while others felt it was a misguided attempt using fear and the threat of punishment to incite change to improve a system that was not as awful as some policymakers claimed it to be (Cochran-Smith, 2005). Simpson et al. (2004) identified NCLB as an "unprecedented…Herculean challenge" (p. 68) in that it was the most onerous and demanding piece of federal education legislation ever enacted to try to reform schools within the U.S. educational system (Albrecht & Joles, 2003; Rentner et al., 2003).

While assessments of NCLB's successes and failures were mixed in the early years after its passing (Cochran-Smith, 2005), overall, there seemed to be more discontent and criticism than anything else. In the few years after NCLB was enacted, Darling-Hammond (2004a) reported that over 20 states and school districts officially opposed the law, and several legislative and educational groups proposed changes that they wanted to occur before the law was up for reauthorization in 2007 (Olson, 2004a).

Additionally, the likelihood of all states meeting all NCLB requirements was found to be slim. Per a report from the Education Commission of the States (2004), while all states were initially on track to meet the NCLB requirements, only 10% were actually likely to meet them all on time. Cochran-Smith (2005) concisely summed up one of NCLB's "fundamental flaw[s]" when she stated that the premise underlying improving the country's educational system was a

> highly coercive accountability system, based on competitive pressures and including public shaming and punishment for failure, will improve schooling for disadvantaged students without the improvement of school capacity, increases in resources, and major investments in programs to improve the quality of professional teachers. (p. 102)

Further, NCLB supported the idea that students' academic achievement—test scores, in this case—accurately depicted what sort of knowledge and abilities educated people should have. Many argued that in addition to content knowledge or abilities, formally educated people should also be positively contributing to society, engaged in democracy, and be good family members and friends. Unfortunately, due to NCLB's onus on test performance, the federal government did not deem any of these virtues as explicitly valuable (Mathis, 2003). In addition, there were several major concerns with each of NCLB's focus areas, and many disputes revolved around fundamental issues in the logic behind the decisions or difficulties in implementing NCLB's new mandates, while other criticisms centered on the fiscal costs of implementing such requirements. These concerns are briefly outlined below.

Student performance and AYP. A big criticism surrounding student performance and AYP was that having the goal of all students (i.e., 100%) meeting the "proficient" standard, even by the 2013-2014 school year, was simply unrealistic (Linn, 2004; Mathis, 2003; Packer, 2004). Linn, Baker, and Betebenner (2002) posited that such a high and unrealistic goal could actually have the opposite effect as intended, in that it could be more demoralizing than motivating for students, teachers, and administrators. In addition to the idea that reaching this goal was likely unrealistic, most states indicated that the most difficult challenge regarding the implementation of practices to meet the NCLB legislation was figuring out how to best assess whether AYP was even being met (Rentner et al., 2003). Further, while AYP was in theory to allow for comparisons among states, the different state standards along with differences in the content or rigor of tests made such comparisons inaccurate or impossible and therefore, essentially useless (Linn, 2004; Linn et al., 2002; Packer, 2004). The added complexity of multiple accountability systems and assessments (e.g., state content standards, AYP, the NAEP) made for additional confusion related to student performance and accountability among education professionals, parents, and the general public (Cochran-Smith, 2005).

Another fundamental issue was the formation of the categories within AYP (e.g., “proficient”). First, using one score to establish the proficiency level (i.e., a cut score) only allowed schools, districts, or states to arbitrarily claim that their students were academically achieving when students made it past such arbitrarily set cut scores. This singular focus, therefore, did not allow for any recognition for improvements in achievement by students who were regularly below or above the cut (Linn et al., 2002). Second, using such a metric at the school level was rife with measurement issues. Linn et al. found that school-level results were especially unreliable from year to year, and this fluctuation and unpredictability could be so great that schools identified within one category (e.g., “proficient”) could be unlikely to be in the same category the following year, simply due to measurement error and other factors (e.g., teacher turnover, student cohort issues) (Kane, Staiger, Grissmer, & Ladd, 2002; Linn & Haug, 2002; Mathis, 2003). Other measurement-related issues included some rural schools or schools with smaller student populations not having enough students in each subgroup to draw valid conclusions in terms of AYP, and some schools with higher levels of poor and/or minority students being labeled as needing improvement because they had more AYP targets to hit (due to their multiple subgroups of students) (Darling-Hammond, 2004a).

Highly qualified teachers. Two dominant issues existed around the HQT provision: 1) knowledge and accountability regarding the HQT requirement and 2) an inequity in the distribution of HQTs. Other criticisms also centered around what constituted a “highly qualified” teacher, difficulties in states successfully meeting the HQT mandate, and whether the HQT provision would truly improve students’ academic achievement.

By the 2004-2005 school year, most teachers were aware of the HQT mandate, however, a significant number were unsure whether they were considered “highly qualified” (due to the multiple ways one could define or become “highly qualified,” depending on a variety of factors). Likewise, many teachers were not notified by their school or district whether or not they met the HQT definition, leading to added uncertainty and confusion (Birman et al., 2007). One possible cause for this uncertainty was that schools and districts had to build completely new data systems to allow for the tracking of a teacher’s HQT status (Rentner et al., 2003), which took time and money to develop and implement. Related, Rebell and Hunter (2004) argued that there were few consequences minimal, if any, enforcement by the federal government for schools or districts that were not meeting the HQT mandate; further, overall, the government appeared to provide scant attention to the tracking of schools' and districts' HQT statuses (Olson, 2004b). The HQT mandate was specifically targeted towards improving the educational outcomes for disadvantaged students. However, three years after NCLB was enacted, while most teachers were deemed as "highly qualified" per NCLB rules, less qualified teachers were more likely to teach disadvantaged and high-needs students (e.g., low-income, minority, special education) compared to "highly qualified" teachers (Birman et al., 2007; Rebell & Hunter, 2004).

Lastly, a big criticism relating to many NCLB mandates, but especially relevant to the HQT provision, was a lack of funding. While Title II of NCLB allocated funds to help recruit, prepare, and train high quality teachers, some argued that the near three billion dollars was simply not enough. Furthermore, and more importantly, researchers noted that no amount of additional funding would automatically increase the teacher supply, which was needed if the HQT goal was to be met at the national level (Simpson et al., 2004).

A new path for NCLB. Although NCLB raised many concerns among government officials, educators, parents, and the public, in 2005, then-Secretary of Education Margaret Spellings announced that there would be a "new path" of "common-sense principles and approaches" that would assist states in meeting NCLB's mandates and goals (U.S. Department of Education, 2008a, para. 1). The main foci of this “common-sense” path were new goals that largely focused on an increase in the frequency of student assessment via standardized tests. With this change, all students between grades three and eight would be tested once per year, and high school students, per grade, would be tested yearly. As before, results would be disaggregated by student subgroups. This change in testing frequency was supposed to result in all students testing at their respective grade level (or higher) (i.e., at “proficient” or better) in both reading and mathematics by 2014 (USDOE, 2008a).

In addition to the increases in the testing of students, this 2005 change also highlighted the use of VAMs at the national level for the first time, which would allow students’ academic achievement to be measured over time. A growth model pilot program for accountability purposes (discussed in more detail, forthcoming) was subsequently announced by the USDOE to see if such models could provide more accurate reports of students’ achievement than the previous NCLB measures, such as AYP. This pilot program further supported the idea of accountability and reform based on student achievement and, related, teacher effectiveness (USDOE, 2008a).

The end of NCLB. Around the same time as the student growth pilot program was ending, *The Widget Effect* (Weisberg et al., 2009) report was released, which further supported the perception that the quality and effectiveness of teachers largely affected student outcomes. In a sample of teachers, Weisberg et al. found that over 90% were classified as “effective” or “highly effective,” yet they reported that it was simply not feasible for so many teachers to be classified as effective or better when the country as a whole was merely performing at an average level compared to similar nations (e.g., Denmark, Finland, Japan). Around the same time this report was released, the American Recovery and Reinvestment Act (ARRA) of 2009—which was supposed to be foundational for improving education reforms—was signed into law, of which the RTTT initiative was a part (discussed in more detail, immediately forthcoming). RTTT, along with the Weisberg et al. (2009) report, helped fuel the push to revamp and reform education accountability policy, especially as related to teacher effectiveness.

Race to the Top

President Obama signed the ARRA into law in 2009, and, as mentioned, a part of this act was the federal RTTT initiative. ARRA funded the RTTT initiative with over $4.35 billion dedicated to states that vowed to make substantial improvements to their education systems. RTTT was a grant competition among states where winners would receive funds to support proposed improvements to their education systems related to teaching and learning. These improvements were supposed to emphasize four main areas: developing "rigorous standards and high quality assessments," attracting and retaining "great teachers and leaders" in the classroom, utilizing "data systems that inform decisions and improve instruction," and supporting and "sustaining education reform" through collaborations among multiple groups to support student achievement and reduce educational gaps (The White House, 2009, para. 5). The first two of the four areas of emphasis—developing standards and high quality assessments and using data systems to inform decisions and improve instruction—became especially important in helping shape the current climate of teacher evaluation in the country (i.e., a focus on standards and data-driven assessments).

RTTT was the first federal initiative or act of federal policy that led to highly consequential personnel decisions for teachers (e.g., being promoted, granted or denied tenure, terminated) being based on their students' test scores, as it called for the use of multiple measures to be a part of states' teacher evaluation systems, as based on student achievement data. RTTT also stipulated that student growth scores not just could but should dictate whether teachers be provided with additional compensation and additional responsibilities, along with whether they received tenure or full certification. While RTTT was positioned as a voluntary competition for states, if states chose to forego involvement they were also choosing to forego federal money that could be used to help improve their educational systems.

RTTT requirements. The grant application and awards were initially structured in two phases, though a third phase was subsequently added (which is discussed in more detail, forthcoming). Specific to the first two phases, states could apply for a grant in either phase, though a state that won an award in Phase I was ineligible to apply for funding in Phase II. Phase I applications were due in January 2010 with awards being announced in March 2010, and Phase II applications were due in June 2010 with awards being announced in August 2010. Funding opportunities ranged from $20 million to $700 million and were dependent on a state's percentage of students out of the national total (Weiss, 2014). For a state to even be eligible to apply for RTTT funds, it had to ensure it did not have any active laws that prevented student achievement or student growth data from being used for teacher or principal evaluations (USDOE, 2009a). This requirement led to six states removing existing laws blocking the usage of such data, and 11 additional states went as far as enacting regulations that actually required such data to be used in evaluations (McGuinn, 2011). Thus, even before RTTT winners were announced or money was disbursed, the competition was already having an impact on educational policy across the nation as nearly three quarters (i.e., 35 of 47, or 74.5%) of all state applicants across both phases ended up updating or creating laws and policies to help them meet and support RTTT demands (USDOE, 2010h).

RTTT application and selection criteria. Grants were awarded based on six selection criteria, each with different weights. Each criterion and its sub-criteria were allotted a certain number of points, and the states that applied and had the highest number of points out of a maximum of 485 points received funds. A heavily weighted criterion was specifically related to teacher evaluation and teacher effectiveness, as of the six criteria, “Great teachers and leaders” (Criterion D) counted for nearly one third of all points. Within that, “Improving teacher and principal effectiveness based on performance” (Criterion D2) made up 42% of Criterion D and 12% of the total possible points, underscoring the importance that the U.S. Department of Education placed on teacher evaluation and performance.

Most notably, RTTT required states to commit to using growth models as a partial way to evaluate teachers and principals (Weiss, 2014), and also called for multiple measures to be used to assess teacher effectiveness. Unlike NCLB, which simply supported the use of such measures, RTTT was the first federal initiative that explicitly required growth measures as a part of teacher evaluation. RTTT also was influential in states adopting the Common Core State Standards (CCSS), as eight percent of a state’s total points relied on the development and adoption of standards based on CCSS (Weiss, 2014). Similar to NCLB, RTTT also had consequences for schools that consistently did not meet expectations. These consequences included options of replacing staff members (including the principal and at least half of all staff members); transferring a school's operational oversight to a charter or educational management organization; transforming several aspects of the school, including replacing the principal, utilizing VAMs as a significant part of the teacher evaluation system, and rewarding or penalizing staff members based on student outcomes; or permanently closing the school (Lohman, 2010).

Grant applications and winners. Forty states plus the District of Columbia (D.C.) applied for RTTT funds in Phase I, and 16 were identified as Phase I finalists based on the tallied numerical scores from their application (USDOE, 2010a). The 16 finalists were invited to present their reform plans to the USDOE, and from that, Delaware and Tennessee were the two Phase I winners receiving approximately $100 million and $500 million, respectively (USDOE, 2010c). Thirty five states and D.C. applied for Phase II grants, and 18 states and D.C. were identified as finalists (USDOE, 2010b). Of those, nine states and D.C. were awarded grants (USDOE, 2010c).

RTTT outcomes. RTTT represented a major shift in U.S. educational policy after NCLB, as the locus of control had shifted from the states back to the federal government. Although the winning states would (try to) implement their own plans, these plans were shaped and governed by the RTTT application stipulations and guidelines, which stemmed from federal oversight (i.e., the USDOE) (Lohman, 2010). Further, RTTT used a strategy opposite of that of NCLB to guide reform, going from using punishment (i.e., sanctioning states that did not meet goals) to rewards (i.e., fiscally awarding states that presented the best reform plans) (McGuinn, 2011; Nee, 2010). Lastly, RTTT provided states with additional funding through a competitive process, whereas previous processes provided federal funding via needs-based formulas (McGuinn, 2014), and, in the case of NCLB, contingent upon compliance with the law (Lohman, 2010).

While many lauded the RTTT initiative as a huge improvement over NCLB and as a policy that would vastly improve the state of American education, especially in light of the recession that began around 2007, the competition was heavily criticized. A major criticism that had drawn out consequences was that winning states ended up overestimating their reform goals and underestimating the time and resources needed to accomplish those goals. Three years after Phase II of RTTT was completed, the majority of states receiving funds were behind schedule in developing and implementing their proposed reform plans, as many RTTT applications contained unrealistic goals and expectations and/or such stated goals were met with unexpected challenges (Weiss, 2014). She noted that all but one winning state had planned to increase student achievement to unreasonable levels, and those levels were not feasible even if the states were to have extra time and/or more money (Weiss, 2014). Further, Boser (2012) asserted that every state needed additional time to implement any of a variety of their reform plans, and specifically, nearly every state needed more time than anticipated to develop its new teacher evaluation systems (Weiss, 2014).

Another major criticism was the use of VAMs in teacher evaluation systems and having high-stakes consequences tied to such output (to be discussed in more detail, forthcoming). As mentioned above, RTTT was the first policy that explicitly required the use of student achievement and VAMs in teacher evaluation systems, and this caused a national uproar as research has found—and continues to find—that such models are rife with measurement and related issues and therefore cannot accurately predict or determine a teacher’s effectiveness (McCaffrey, Sass, Lockwood, & Mihaly, 2009; to be discussed in more detail, forthcoming).

On a more neutral note, as previously mentioned, one of the immediate outcomes of RTTT was an uptick in the amount of education related policies that were amended or enacted for the first time. In line with the USDOE (2010h), Howell (2015) noted that simply the process of many states applying for RTTT funds led to large increases across the country in the percentages of proposed education reform policies that actually became signed into law (e.g., those related to high quality or rigorous standards, alternative paths to teacher certification, measuring student growth, support for charter schools). He noted that this policy increase occurred even in states that did not apply for any RTTT funds, and ascribed the increase to the overall discussion and fervor that the RTTT competition created across the nation.

The implications of RTTT were far-reaching, as it cemented and continued the increased fervor regarding all aspects of school reform, but especially test-based accountability, creating charter schools, and developing and implementing common standards (McGuinn, 2014). McGuinn also noted that RTTT stimulated philanthropically-motivated foundations, political organizations, and entrepreneurial educators with interests in and ties to education reform (e.g., Bill and Melinda Gates

Foundation, Mark Zuckerberg, Teach for America) to increase their roles and influences in the country's educational reform goals and, at times, to further enhance RTTT efforts.

RTTT Phase III and NCLB Waivers

The end of the 2000s and the beginning of the 2010s saw continued efforts to transform and revamp the U.S. education system with specific standards-based and accountability reforms. In 2011, then-President Obama officially requested an additional $1.35 billion to support a third round of the RTTT competition as ARRA money had been fully allocated and the USDOE, especially then-Secretary Arne Duncan, lauded the RTTT competition and its outcomes (see, for example, McGuinn, 2014). Much less than Obama's request, the government provided approximately $200 million for a third phase of RTTT as a part of the Fiscal Year 2011 Appropriations Act (USDOE, 2011a). Also occurring in 2011 was the creation of NCLB "waivers" (described in more detail, forthcoming), which allowed states to bypass certain NCLB mandates. These waivers stemmed from the USDOE's (2012b) realization and admittance that NCLB unintentionally had several major flaws, including allowing or even encouraging states to set low standards, failing to both recognize and/or reward student growth achievement, and failing to recognize effective teachers. The following subsections briefly describe Phase III of RTTT and the NCLB waivers.

RTTT Phase III. The USDOE (2011d) released its call for Phase III applications at the end of November 2011, with applications being due in mid-December 2011. RTTT Phase III differed from the previous two phases in that only the nine Phase II finalists that did not win a grant were eligible to apply. To apply for a Phase III grant, as for Phase II, a state still had to have no barriers to using student achievement or VAMs to evaluate teachers and principals; had to be committed to improving its assessments, which had to be aligned with a common set of standards; and had to still support and commit to all proposed reforms in its Phase II application (USDOE, 2011d). Phase III followed the same premises and priorities of Phases I and II, though Phase III had a stronger focus on improving STEM education. Of the nine eligible states, seven applied and ultimately received shares of the $200 million award.

NCLB waivers. NCLB was originally supposed to be up for reauthorization in 2007; however, the reauthorization never occurred due to political conflict in Congress and, therefore, it remained in effect as originally enacted (McGuinn, 2014). To remedy this situation, beginning in 2011, states were allowed to apply for "flexibility," or NCLB waivers, which would allow states to bypass specific NCLB mandates provided states had "rigorous and comprehensive…plans designed to improve educational outcomes for all students, close achievement gaps, increase equity, and improve the quality of instruction" (USDOE, 2012b, p. 1). Waivers would be effective through the 2013-2014 academic year, and be eligible for a one-year renewal. Waivers were allowed for all aspects of NCLB, including school and district improvement requirements, teacher quality and teacher and principal evaluation requirements, and funding and grants. While the NCLB waivers were voluntary, most states completed the application process. If states so chose to request flexibility in their teacher evaluation systems, they were required to create and utilize such systems that, in part, had at least three different performance levels (e.g., ineffective, effective, highly effective) and used VAMs as a significant factor in evaluating teachers and school quality (USDOE, 2012b).

In particular, the waivers allowed states to request exemption from the NCLB mandate that required 100% AYP proficiency for all student subgroups by the 2013-2014 school year. States were instead allowed to adjust their annual measurable objectives to one of three options: reducing the achievement gap between at-risk students and non-at-risk students by 50% within six years, all subgroups achieving 100% proficiency by the 2019-2020 school year, or a different state-designed plan that was just as rigorous and challenging as a 50% reduction in the achievement gap (USDOE, 2012b; see also McNeil, 2012).

A major component of the NCLB waivers as related to accountability and teacher evaluation was the requirement that states use VAMs in their teacher evaluation systems. With states no longer utilizing AYP, a student-based measure of success, the landscape and onus of accountability had fully shifted to teacher-based accountability. While the RTTT competition had previously resulted in many states revamping their teacher evaluation systems to include such data, the NCLB waivers cemented this practice as waiver requirements stipulated that multiple measures be used to evaluate teachers and principals, including student growth "as a significant factor" (USDOE, 2012b, p. 51). Waiver guidelines stipulated that VAMs not just simply be used in teacher evaluation systems but that it (along with other evaluation measures, such as classroom observations, teacher portfolios, student and/or parent surveys, etc.) be used to inform personnel decisions. While the guidelines suggested that states and districts avoid firing any teachers solely based on a single student growth score, states and districts were not expressly prohibited from doing so (USDOE, 2012b). A further component of the waivers, which is not discussed in detail here, was the push for states and districts to support its highly effective teachers with benefits such as additional or performance-based payments, among other non-fiscal rewards. This focus, especially on additional compensation, further increased the intensity of teacher-level accountability mechanisms.

The Every Student Succeeds Act (ESSA)

Even with the NCLB waivers in place, ESEA had still yet to be reauthorized, which it had been up for since 2007. Finally, in late 2015, then-President Obama signed the bipartisan-supported Every Student Succeeds Act (ESSA, 2015) into law. ESSA was especially noteworthy not just because it was eight years overdue, but also because it represented the first time in decades where the federal government's role in education policy was reduced and more control was given back to the states.

As mentioned, the government realized that the one-size-fits-all, top-down approach (e.g., used with NCLB) did not work (USDOE, 2012b). ESSA provided states with the opportunity to redesign their school accountability mechanisms and teacher evaluation systems, giving states more flexibility than they had with NCLB. However, the focus of ESSA still remained on accountability, as well as providing a "high quality" education for all students, supporting equitable outcomes, and protecting students' civil rights, especially for those students who identify with historically marginalized groups.

State ESSA plans. Once ESSA was passed, since accountability was no longer in the hands of the federal government but under the purview of individual states, each state was required to submit to the federal government a plan that detailed how it would hold schools accountable, measure student success, and evaluate teachers (among other requirements, though those are out of the scope of this study and thus are not detailed here), with the overarching goal being improving student outcomes (Aldeman, Hyslop, Marchitello, Schiess, & Pennington, 2017). States were required to submit their plans in either April or September of 2017 (with the goal of implementing the plans in the 2017-2018 or 2018-2019 school years), after which the plans were then reviewed by the Department of Education, along with teams of parents, teachers, principals, other school leaders, community members, and researchers. The purpose of the review was to "maximize collaboration with each [s]tate," "promote effective implementation of challenging…academic standards through…innovation," and "provide transparent, timely, and objective feedback…designed to strengthen the technical and overall quality" of each state's plan (USDOE, 2017e, p. 3). Sixteen states and the District of Columbia submitted their plans in April 2017, and the remaining states in September 2017 (USDOE, 2017b).

School accountability and student success. Through ESSA, states were encouraged to develop accountability systems that best meet their respective needs. States were allowed to incorporate new and additional measures into their accountability formulas to assess school and student success, though they had to remain committed to working collaboratively with their schools and districts to ensure success for all students, including those from individual subgroups, and to "turn around" poorly performing schools (USDOE, 2016b). ESSA required states to measure progress on four indicators:

academic achievement, academic progress (for elementary/secondary schools that do not award diplomas) or graduation rates (for high schools that do award diplomas), progress in achieving English language proficiency, and school quality or student success (USDOE, 2017a). The last indicator, school quality or student success, did not need to be academic in nature and, for the first time in federal policy history, could include more latent concepts like school climate or student engagement.

To determine school success, states had the flexibility to create and/or select what measures or indicators they would use, along with determining the relative weights of each indicator (though academic factors like test scores and graduation rates had to count more heavily than other non-academic factors), with the only requirements being that all measures must be used for all public elementary and secondary schools in the state, including charters; be valid and reliable; allow for the comparison of subgroups of students; and measure several specific outcomes, including academic achievement and graduation rates/student progress (USDOE, 2016b). Importantly, ESSA allowed states to utilize measures of student success that identified the potential progress of *all* students, rather than using one measure with one cut score (i.e., AYP from NCLB) that excluded students who were consistently above or below the cut (Linn et al., 2002).

While ESSA greatly increased the flexibility that states had in choosing accountability- and progress-related measures, it still emphasized standardized testing, as it required on an annual basis all students in grades three to eight and once in high school participate in statewide reading and mathematics assessments. The logic was that such annual measures would allow for “a fair and accurate picture of school success,” as well as to help determine the achievement of student subgroups (USDOE, 2016b, p. 3). However, states could choose exactly when and how the tests were administered; set a limit on the amount of time students must spend taking tests; use provided funds to determine what assessments, if any, are unneeded and can therefore be removed from the testing regimen; and create policies that allow parents/guardians to opt their children out of such testing (USDOE, 2016b; see also Walker, 2015).

From a school quality standpoint, the previous signaling labels from NCLB of schools being “satisfactory” and “in needs of improvement” schools were removed with ESSA. For poorly performing schools, ESSA regulations required that states identify schools by the 2018-2019 school year that needed additional improvement and support, along with identifying schools with “consistently underperforming subgroups” on a yearly basis starting in the 2019-2020 school year (USDOE, 2016b, p. 2). Every three years, states had to also identify schools needing “comprehensive support and improvement,” or the lowest five percent of Title I schools based on performance, high schools with graduation rates under 67%, and Title I schools with consistently poorly performing subgroups who have not improved after targeted improvement plans (USDOE, 2016a). In keeping in line with flexibility, states were allowed to create their own definitions of what “poorly performing” entailed, which is another deviation from the more structured regulations of NCLB.

Teacher evaluation. The design and implementation of teacher evaluation systems, and thus, how teachers are evaluated, had the potential to undergo great changes under ESSA. One of the most significant aspects related to teacher evaluation under

ESSA was the transition of authority from the federal government to the individual states, and in many cases, from the states to local districts (Desimone et al., 2019). Along with this transition was the call for teacher evaluation systems to shift from primarily focusing on accountability and thus personnel decisions to providing more useful formative feedback to teachers that would lead to their improved personal growth (Connally & Tooley, 2016). Under ESSA, no longer could the federal government dictate how teachers are to be evaluated, nor can the government require that VAMs must be used in teacher or principal evaluations. Growth models *could* be used, but their use was no longer a mandatory component of teacher evaluation systems. Rather, ESSA supported changes to teacher evaluation systems being developed collaboratively among teachers, administrators, parents, and other stakeholders (USDOE, 2016b; see also Walker, 2015).

Current landscape of teacher evaluation systems under ESSA. Around a year after all states submitted their initial ESSA plans to the federal government, many states' teacher evaluation systems seemed, on the whole, largely unchanged from their pre-ESSA days (Close et al., 2018). However, specific to VAMs, Close et al. noted that the number of states using VAMs or encouraging their use had declined since ESSA was passed. Only approximately 30% of states still were in support of or encouraged districts to use VAMs, while around 45% of states explicitly no longer encouraged VAM use. While the decline in VAM use and support was promising, given VAMs' noted measurement and pragmatic concerns (discussed in more detail in the VAM Research section, forthcoming), the decline was not as immediate as initially hoped or anticipated (Close, Amrein-Beardsley, & Collins, 2019; Loewus, 2017) as many teachers, administrators, researchers, and community members called for more states to actively discourage using VAMs or outlaw them.

An additional change under ESSA was the increased flexibility of how “student growth” was defined and how VAMs were utilized within teacher evaluation systems. Under ESSA, states and/or districts could decide, for example, how much weight VAMs should carry in a teacher’s overall evaluation score, whether VAM results are to be used for summative and/or formative purposes, and which measures of teacher effectiveness could be used in determining a teacher’s final yearly summative evaluation rating (Close et al., 2018). This was a huge divergence from NCLB and RTTT, both of which had stipulated many uniform mandates across all states and/or districts. Additionally, one of the biggest departures from NCLB and RTTT with ESSA was how states planned on using teacher evaluation measures. Prior to ESSA, many summative teacher evaluation ratings, which were largely informed by teachers’ VAM scores, were being used to make or justify highly consequential personnel decisions (e.g., those involving teacher termination, contract renewal, merit pay). Under states’ ESSA plans, teacher evaluation measures appeared to be more formative in nature, with the goal of teacher evaluation being to provide teachers with actionable feedback that allowed them to hone their skills (and thus improve student outcomes), rather than to justify personnel decisions (Close et al., 2018).

Outside of VAMs, states continued to utilize a “multiple measures” approach to teacher evaluation, where multiple indicators are to be used to evaluate a teacher’s performance. Many states (i.e., over 70%; Close et al., 2019) continued with their prior use of classroom observations, while some also included new(er) measures as well like student and/or parent surveys, portfolios, student learning objectives (SLOs), and more.

At the time of this writing, while the consequences of states' and districts' updated teacher evaluation systems are yet to be fully realized, it appears that, at the very least, local control of a variety of aspects of teacher evaluation is at an all-time high. As Close et al. (2019) noted, many states found this increased local control to be extremely valuable as people integral to the teacher evaluation process (e.g., teachers, administrators) finally have an opportunity to be heard.

To date, the most significant impact of this local control regarding teacher evaluation has resulted in the reduction of test-based accountability requirements (e.g., standardized tests, VAMs; Desimone et al., 2019). While ESSA has been the guiding policy for school and teacher accountability for several years already, questions still remain about its outcomes as recent reports have indicated that many local schools and districts have yet to truly see or feels its effects (Klein, 2019).

History of Teacher Evaluation in New Mexico

Education reform was deemed to have officially begun in New Mexico in 2003 (New Mexico Public Education Department [NMPED], 2010a), via House Bill (HB) 212 (New Mexico Legislature, 2003). HB 212 created high academic standards, a three-tier licensure program for teachers, an improved data system, and an online database for K-20 curriculum (NMPED, 2010a). The provisions of HB 212 were in direct alignment with NCLB's reform areas (e.g., assessments, student achievement, and AYP; data systems)

(New Mexico Legislature, 2003), and therefore, the bill did not specifically include provisions about teacher evaluation.

With HB 212 in place, several years after both NCLB and HB 212 were enacted, only 50% of schools in New Mexico were meeting the federal AYP target, while the average across all states was 70%. Further, over half of the schools in New Mexico were either designated as “in needs of improvement” or, worse, needing to be restructured, compared to just 13% across the country. Further, at the time, compared to the national average, New Mexico was behind in reading and mathematics achievement, high school graduation rates, and the number of students who took Advanced Placement (AP) exams (USDOE, 2008b). Additionally, in the year immediately following NCLB, the number of schools in New Mexico meeting AYP dropped by 21 percentage points between the 2003-2004 and 2005-2006 school years (Taylor, Stecher, O’Day, Naftel, & Le Floch, 2010). While NCLB was supposed to improve outcomes for students across all states, and HB 212 appeared to create plans to further support those improvements in New Mexico, those improvements were never realized.

The School Personnel Act of 2006

While HB 212 marked the beginning of education reform in New Mexico, the School Personnel Act of 2006 was the first statute that solidified uniform statewide teacher evaluation standards in the state. Per this legislation, teachers were evaluated in part via classroom observations conducted by their principals, who would determine whether a teacher could adequately demonstrate the agreed upon statewide competencies. The frequency of observations was based on the level of each teacher (i.e., Levels I, II, or

III), with Level I teachers being evaluated yearly and Levels II and III teachers being evaluated every third year (USDOE, 2011b). The other component to teachers' evaluations were how well teachers carried out professional development plans they created in conjunction with their school principal at the beginning of each school year. This teacher evaluation system was a binary system, in that a teacher could receive one of only two performance ratings: "effective/meets competencies" or "ineffective/does not meet competencies." This lack of differentiation led many teachers to be labeled as "effective," which some (e.g., Weisberg et al., 2009), including then-Secretary of Education Arne Duncan (Heinz, 2011), thought was highly problematic given New Mexico's students' subpar student achievement.

Per the School Personnel Act (2006), regarding teacher termination, if a teacher's performance was deemed to be less than satisfactory, the school principal could require the teacher to undergo "peer intervention" (Part D), which could include mentoring, for a length of time determined by the principal. Further details about what peer intervention entailed were not specified in the act. If a teacher could not improve his/her performance over the time period specified by the principal, the teacher could be recommended for termination, on which the local school board would ultimately decide. The teacher evaluation system at this time was, overall, quite subjective as the school principal made the majority of evaluation- and termination-related decisions (e.g., conducting the observations, specifying peer intervention requirements, recommending teachers for termination). When the School Personnel Act (2006) was governing the state's teacher evaluation system, student growth data (i.e., VAMs) was not a part of the state's teacher evaluation system.

Race to the Top in New Mexico

New Mexico submitted its Phase I RTTT application in January 2010 and emphasized that its proposal addressed and aligned with all six of the RTTT priorities (see NMPED, 2010a, pp. 9-10). The RTTT application highlighted the state's current successes (e.g., establishing a uniform teacher evaluation system), but indicated that further achievements, such as "higher levels of student achievement and success," (p. 11) could not be met without additional fiscal assistance, such as that from the RTTT grant. The state received 325.2 points out of a possible 500 (i.e., 65% of the total possible points), and reviewers noted several concerns regarding the potential implementation of New Mexico's proposed initiatives. Such concerns included many small rural districts in the state would be hard to reach with reform efforts, the application's absence of a discussion of how the state would utilize its own current fiscal and personnel resources to support RTTT efforts, misleading application information about student achievement progress, and a lack of support from the Albuquerque Teachers Federation (USDOE, 2010d, 2010g). Of the 41 RTTT Phase I applicants, New Mexico ranked 30th and was not identified as a Phase I finalist (USDOE, n.d.a).

New Mexico tried again to win a RTTT grant in Phase II of the RTTT competition, as it submitted its Phase II application in May 2010 (NMPED, 2010b). It improved upon its Phase I application by addressing previous areas that were lacking and/or concerning to reviewers. Its Phase II application received 366.2 points out of a possible 500 (i.e., 73% of the total possible points), which ranked them as 28th among the 36 Phase II applicants, and, again, New Mexico was not identified as a RTTT finalist (USDOE, n.d.b, 2010e). Regarding the state's Phase II application, reviewers noted that, overall, the Phase II application contained goals that were overly ambitious and/or not compelling enough to warrant RTTT funds, robust explanations of how certain interventions and initiatives would be successful was lacking, there was a lack of support from the state's Native American leaders, and there appeared to be confusion over how the state's current funds would be used to support reform efforts (USDOE, 2010f). As New Mexico was not a Phase II finalist, it was not eligible to apply for Phase III funds in 2011 (USDOE, 2011d).

SB 502 and Executive Order 2011-024

In January 2011, a new governor, Susana Martinez, took office and she, in conjunction with the NMPED, was committed to the new "Kids First, New Mexico Wins" education reform plan. This plan focused on increasing school and teacher accountability, aiding struggling students and schools, prioritizing education funding to directly help students, and reward effective teachers and leaders (NMPED, 2011). Through this plan, Martinez vowed to make improving education a priority for the state.

In February 2011, Senate Bill (SB) 427—A-B-C-D-F School Rating System—and SB 502—School Teacher and Principal Evaluation—were proposed, as the state was committed to education reform even though it failed twice to secure RTTT funds. SB 427, which was signed into law in March 2011, led to the creation of an A through F letter grading system for schools for accountability and transparency purposes. Letter grades were determined in part by student achievement on standardized tests and reading and mathematics student growth, both overall and within the lowest quartile of students. While SB 427 was not directly tied to teacher evaluation, it was one of Martinez's first reform efforts to improve accountability in the state.

Unlike SB 427, SB 502 was ultimately not passed, though it was instrumental in the formulation of the "outputs-based" (Paul, 2015) teacher evaluation system that was called into question specifically in the 2015 lawsuit (*State ex rel. Stewart v. New Mexico Public Education Department*, 2015). Among SB 502's components was the requirement that schools adopt a new teacher evaluation system to improve student achievement by "identify[ing] teachers who are most effective at helping students succeed" (p. 2). SB 502 stipulated that this teacher evaluation system should be comprised of multiple methods, including student growth data (i.e., VAMs); classroom observations; and other measures, such as student and/or parent surveys, peer observations, teacher portfolios, or other approved options. Specific to VAMs, SB 502 stipulated that student growth data be the highest weighted component of evaluations for teachers in tested subjects and grades. SB 502 also required the creation of a work group to develop and recommend the new statewide teacher evaluation system, as well as proposing a performance pay system to incentivize teachers and principals. Further, and perhaps the most consequential, SB 502 proposed that any teacher who earned the lowest effectiveness teacher evaluation rating for three years in a row would be fired, unless the teacher could demonstrate that his/her evaluation results were inaccurate. The goal of SB 502 was to have the new evaluation and pay systems in place by the 2013-2014 school year.

There were several notable issues with SB 502, including concerns about different districts developing different evaluation and compensation plans, possible legal issues that could stem from "unproven…evaluation methods" (Gudgel, 2011, p. 6) that would be used to terminate teachers, and a substantial increased cost to the state (Gudgel, 2011). SB 502 also was in direct conflict with other established bills, including the aforementioned A-B-C-D-F School Rating System (SB 427), the School Personnel Evaluation System of 2011 (SB 503), and the Teacher Choice Compensation Fund of 2011 (SB 567). Thus, the bill did not pass.

Although SB 502 did not pass, Governor Martinez remained committed to making the state's education a priority by creating the New Mexico Effective Teaching Task Force (NMETTF) via Executive Order 2011-024 to help recruit, retain, and reward the state's best teachers (State of New Mexico, 2011). As previously noted, the then-current teacher evaluation system was quite subjective (i.e., outcomes were mainly determined by the school principal) and also emphasized years of experience and credentials, but did not include any component of student growth data (Office of the Governor, 2011). This lack of inclusion of student growth data was seen as detrimental to both teacher retention and student achievement.

The task force was comprised of 15 members "with over 100 years of [total] classroom experience" (Office of the Governor, 2011, para. 1) who met multiple times over the summer in 2011 to: (a) recommend measures of student achievement to best evaluate teacher performance, (b) identify best practices of effective teachers and teaching to compose the remainder (i.e., other than student growth) of teacher performance evaluations, (c) recommend appropriate weights for (a) and (b), and (d) identify how New Mexico could convert to a performance pay system based on student growth data (NMETTF, 2011).

Of the 38 recommendations from NMETTF (2011), the first eight were directly related to teacher evaluation (with another four specific to school leader [i.e., principal] evaluation). The eight teacher evaluation related recommendations from NMETFF were:

> (1) replacing the binary pass/fail teacher evaluation system with five effectiveness categories;
>
> (2) assigning one of the five effectiveness categories to teachers only after a "careful consideration" (p. 5) of multiple inputs (e.g., student growth data, observations);
>
> (3a) using a VAM to "reliably capture student achievement", as well as disseminate individual VAM scores to the teachers to "inform instruction" (p. 5);
>
> (3b) for teachers in tested grades and subjects, weighting their evaluation as follows: 50% based on the teacher's VAM score, 25% based on classroom observations, and 25% based on other multiple measures that have been approved by the PED and locally adopted;
>
> (4a) implementing the use of VAMs in phases, first for teachers in tested subjects and grades and then for teachers in non-tested subjects and grades, and implementing the evaluation via observations and other multiple measures immediately;
>
> (4b) for teachers in non-tested subjects and grades, weighting their evaluation as

follows: 25% based on the teacher's school's A-F letter grade, 25% based on classroom observations, and 50% based on other multiple measures that have been approved by the PED and locally adopted;

(5) continuing to use classroom observations with objective protocols, in addition to VAMs, for evaluative purposes;

(6) allowing key stakeholders, such as teachers, principals, parents, etc. to provide input on teacher evaluation policies;

(7) utilizing a matrix where all components of the teacher evaluation system are combined to produce an overall effectiveness rating; and

(8) providing a post-evaluation conference with each teacher with "actionable feedback" (p. 6), though not using this conference to terminate a teacher.

Several months after the NMETTF submitted its final report to the state, then-President Obama announced the plan to allow states to apply for ESEA flexibility, (i.e., the NCLB waivers). In New Mexico, then-Secretary of Education designee, Hanna Skandera, deemed that in order for New Mexico to be competitive for a NCLB waiver, the state had to implement the NMETTF's (2011) recommendations (Heinz, 2011). Thus, New Mexico's new teacher evaluation system was born. This system is described in detail in the following section.

The NCLB Waiver and the New Teacher Evaluation System

In November 2011, New Mexico applied for a NCLB waiver regarding NCLB's AYP requirements, use of federal funds, and a development plan regarding HQTs (USDOE, 2011c). Regarding accountability, the waiver request asked for permission to use the newly instituted A-F grading system instead of the (binary) AYP system for school accountability purposes (USDOE, 2011c, 2012c). Although the waiver was initially declined due to several concerns, one of which was a lack of identifying student subgroups for accountability purposes (USDOE, 2011b), the USDOE granted an updated waiver request in early 2012, which made New Mexico the 11th state to have received a NCLB waiver (USDOE, 2012a).

When New Mexico had submitted both its original and then updated NCLB waiver request, the state had yet to adopt a new and improved teacher evaluation system. Thus, as a part of the conditions of the waiver (see USDOE, 2011c), the state had to create the foundation for such an evaluation system prior to the 2012-2013 school year. Using recommendations from the NMETTF (2011), HB 249—the Teacher and School Leader Effectiveness Act of 2012—was proposed to solidify the state's new teacher evaluation system, which included many of the NMETTF's (2011) recommendations. Briefly, HB 249 proposed to get rid of the previous binary teacher evaluation classifications of "meets competencies" or "does not meets competencies" and replace them with the five ratings suggested by the NMETTF. The bill also stipulated that 50% of teachers' evaluations would be comprised of student growth data; there would be a minimum of four classroom observations per year, regardless of a teacher's level or license; and other PED-approved multiple measures would be used in addition to VAMs and observations. The HB 249 proposal did not require the actual immediate implementation of this new evaluation system (which was in line with the state's NCLB waiver), but instead laid the groundwork (i.e., a framework) for such an evaluation system to be put into future use. Per the state's NCLB waiver application, the state planned on beginning the initial implementation stages of this new teacher evaluation system by the 2013-2014 school year and to have it fully implemented by the 2014-2015 school year (USDOE, 2011c).

Although HB 249 ultimately did not pass due to the legislature's adjournment for the year, in April 2012, Governor Martinez instructed the NMPED—which notably had such regulatory authority—to move forward with implementing this new statewide evaluation system. As a part of these implementation plans, NMPED established the New Mexico Teacher Evaluation Advisory Council (NMTEACH) to identify best ways for the state to implement the new evaluation system (USDOE, 2012d). In August 2012, NMAC 6.69.8 (2012) officially established the new teacher evaluation system, which later became known as the NMTEACH Educator Effectiveness System (NMTEACH EES). Per the NMTEACH EES, beginning in the 2013-2014 school year, teachers' evaluations would be comprised of 50% student growth data (i.e., VAM scores), 25% classroom observations, and 25% of other PED-approved measures. Additionally, the previous binary ratings of "effective/meet competencies" or "ineffective/does not meet competencies" were replaced by five levels of performance ranging from "ineffective, does not meet competency" to "exemplary, meets competency."

With a teacher's VAM score counting for 50% of their evaluation, both classroom observations and the other measures utilized (e.g., teacher attendance, student or parent surveys) counted for varying percentages based on how many years of student growth data a teacher had. The NMTEACH EES dictated that all teachers be observed in their classrooms annually, regardless of their tenure status. Teachers had to be observed either three times by the same observer, or twice by two different observers. All observers had to undergo formal training and certification, as well as previously had to have received an “effective” teacher rating or higher (i.e., “highly effective,” “exemplary”). The state also required that all districts use a centralized database to house teacher evaluation data. This database ensured that student achievement data was correctly linked to students’ respective teachers, and also allowed for multiple teachers to be associated with a single student.

At this time, teacher evaluation scores were not considered in a teacher’s tenure review; rather, tenure was a near-automatic process and was granted after a teacher’s three-year probationary period with no regard to his or her effectiveness classification. Conversely, while effectiveness was not considered regarding tenure, effectiveness was a major consideration in the decision to terminate teachers. At the time, New Mexico was one of 28 states where teachers being classified as either “ineffective” or “minimally effective” was grounds for dismissal. A teacher who was no longer on probation (i.e., the teacher had been employed for more than three years) and who was rated as “ineffective” or “minimally effective” was first placed on a three-month performance improvement plan. If the teacher did not make demonstrable progress after those 90 days, the school superintendent would then determine if the teacher should be terminated. If terminated, teachers had two opportunities to appeal the initial termination decision, the second of which was decided by an arbitrator. However, if that the termination appeal went to arbitration, the arbitrator’s decision was final (National Council on Teacher Quality [NCTQ], 2015).

While many across the state lauded the efforts put into the new teacher evaluation system, the NMTEACH EES ultimately would be the source of much consternation and debate for years to come. The first of three lawsuits related to education and teacher evaluation policy was filed a year after the NMTEACH EES was created, with two additional different lawsuits filed in each subsequent year (i.e., one in 2014 and one in 2015). These lawsuits, and the state’s NCLB waiver renewal process, are detailed in the following sections.

The First Lawsuit

In September 2013, the New Mexico teachers’ union group, the American Federation of Teachers New Mexico (AFTNM), sued the NMPED and then-Secretary of Education designee Skandera on the grounds that Skandera had abused her executive power by setting policy related to teacher and school leader evaluation (*State ex rel. Stapleton v. Skandera*, 2013). The AFTNM claimed that Skandera overstepped her authority to enact NMAC 6.69.8 (2012), and that two provisions within NMAC 6.69.8 violated the state’s Public School Code. A state court judge and an appeals court, in 2013 and 2015, respectively, ultimately threw out the suit as both the state court judge and the appeals court found that Skandera did not exceed her authority as Secretary of Education designee, nor did any provisions of NMAC 6.69.8 (2012) violate the state’s Public School Code.

The NCLB Waiver Renewal

New Mexico’s NCLB waiver was initially set to expire at the end of the 2013-2014 school year, and the USDOE (2014b) offered one-year flexibility extensions to any state that was previously approved in Windows 1 or 2 of the initial NCLB waiver process, of which New Mexico was. New Mexico applied for and was granted the one-year extension due to its reported and documented successes with its initial waiver implementation (USDOE, 2014c).

After the one-year extension opportunity, all states that had previously received NCLB waivers—regardless of whether they applied for and/or received one-year waiver extensions—were offered the opportunity to apply for three- or four-year waiver renewals, depending on their initial ESEA flexibility commitments. Renewal requests were due by the end of March 2015, though states (e.g., New Mexico) that were awarded a waiver in Window 1 or 2 and were fully meetings their commitments were invited to submit a renewal request by the end of January 2015 for an expedited review. A state’s renewal request needed to a) document how initial waivers were effective, b) detail how it would continue to meets its initial flexibility commitments, and c) amend any previous waiver requests, if wanted. To receive a three- or four-year waiver renewal, a state also had to outline how it resolved any previously outstanding issues related to implementing its initial flexibility (USDOE, 2014a).

New Mexico was one of seven states that were invited to submit a renewal with an expedited review, and in March 2015, its renewal was approved through the 2018-2019 school year (USDOE, 2015c). The state’s education reform policies were seen as highly successful and moving in the right direction to due to the focus on using student growth data for teacher evaluations and school grades, transitioning away from its binary (i.e., pass/fail) school and teacher accountability system to one that provided “actionable information about…performance” (USDOE, 2015c, para. 12), and its initiative to enroll more students in Advanced Placement (AP) exams (Swedien, 2015; USDOE, 2015c). Under the updated flexibility, New Mexico’s previous approved waivers still applied. While the changes made under the waiver renewal were not specific to teacher evaluation, they were related to overall accountability (e.g., identifying and rating Title I schools, transitioning to the Partnership for assessment of Readiness for College and Careers [PARCC] assessments; USDOE, 2015a). Specific to teachers, with NMTEACH EES now in place, teachers who received ratings of “effective,” “highly effective,” or “exemplary” would be classified as HQTs and the state would need to make public a report of the numbers of HQTs employed in each school and district (USDOE, 2015b).

The Second and Third Lawsuits

One year later after the first lawsuit was filed (*State ex rel. Stapleton v. Skandera,* 2013), the National Education Association New Mexico (NEANM)—an advocacy group for New Mexico public schools, staff members, and students (NEANM, 2018)—sued Skandera on the grounds that the local control that school districts were supposed to have over teacher evaluations was taken away by the very policy that Skandera had used her power to update (i.e., NMAC 6.69.8, 2012). In the suit, the NEANM asked for the state’s teacher evaluation system as governed by NMAC 6.69.8 to be declared illegal and prevented from further use (*State ex rel. Stewart v. Skandera*, 2014). The NMPED requested that the court provide a judgment on the suit, though the court ultimately denied this request and the case headed to discovery. Progress on the case stalled, and in late November 2018, a Notice of Inactivity was placed on the case. The case was subsequently dismissed via an Order of Administrative Closure at the end of May 2019.

As previously described (see Chapter 1), in early 2015, the AFTNM, along with the Albuquerque Teachers Federation (ATF) and several teachers and politicians filed suit against the NMPED and Skandera yet again, claiming that the NMTEACH EES was detrimental to both teachers and students and that it violated the state's "highly objective and uniform" evaluation requirement for all teachers (*State ex rel. Stewart v. New Mexico Public Education Department*, 2015; see also NMAC, 2011; AFT, 2015). As noted in Chapter 1, the presiding judge granted a preliminary injunction that prevented the NMTEACH EES from being used to make any further consequential decisions for teachers until the state could provide evidence that the NMTEACH EES was reliable, valid, unbiased, and fair.

Around this time, the plaintiffs' lawyers called upon Dr. Audrey Amrein-Beardsley to serve as an expert witness on the case, given her expertise in VAMs, teacher evaluation policy, and related accountability policy. She was tasked with analyzing the state's teacher evaluation data (i.e., VAS scores; classroom observation scores; Planning, Preparation, and Professionalism [PPP] scores; and student perception survey [SPS] scores) from the 2013-2014, 2014-2015, and 2015-2016 school years regarding the measures' levels of statistical reliability, validity, and bias (or lack thereof), with a specific focus on the VAS data given it was the most heavily weighted component of the teacher evaluation system and also given its notable contentious attributes (these measurement concepts are discussed in more detail in the VAM Research and Conceptual Framework sections, forthcoming).

Analyses for *State ex rel. Stewart v. New Mexico Public Education Department* (2015). Results from Amrein-Beardsley and Geiger's (revise and resubmit) analyses of New Mexico's teacher evaluation system during the 2013-2014 through 2015-2016 school years indicated that, overall, the system was unfair as only approximately 30% of teachers across the state were eligible to be assessed via VAMs, per year.

Specifically, regarding reliability (i.e., consistency over time), nearly 40% of all VAM-eligible teachers differed in their effectiveness ratings by one quintile and approximately 28% differed by two or more quintiles from year to year. Regarding convergent-related validity (i.e., the relationship between two measures that theoretically assess the same underlying construct), correlation values between teacher's VAS scores and classroom observation scores were notably weak, ranging from $r = 0.15$ to $r = 0.21$.

Lastly, in terms of bias (i.e., scores systematically varying for certain groups of teachers based factors that are not relevant to the scores themselves), several groups of teachers had statistically significantly lower VAS scores than other teachers. Teachers with lower VAS scores included those with fewer years of experience (compared to teachers with more years of experience), those who taught primarily SE or ELL students (compared to teachers who did not teach such students), and those who taught in schools with high relative proportions of SE, ELL, FRL, and URM students (compared to teachers who taught in schools with lower relative proportions of such students). These findings, especially as situated within the related research literature (as discussed in more detail, forthcoming), implied that the system and its measures were, broadly speaking, unreliable, invalid, and likely biased.

Between 2015 and 2018, multiple depositions and hearings occurred, including a deposition where Dr. Amrein-Beardsley presented and discussed the above findings. Ultimately, while the then-current evaluation system remained in place, the preliminary injunction also remained as at no point was the state able to demonstrate that the teacher evaluation system and its measures were reliable, valid, unbiased, and fair.

In early 2019, newly elected Democratic governor Michelle Lujan Grisham was sworn in to office, and only two days later signed an Executive Order (State of New Mexico, 2019) that amended the state's teacher evaluation system. The Executive Order mandated that teachers would no longer be evaluated by VAMs, and that the NMPED must work with a group of stakeholders to "determine more appropriate methods of measuring teacher efficacy and performance" (p. 2). With this Executive Order in place, the plaintiffs decided to no longer pursue the lawsuit (see Amrein-Beardsley & Geiger, 2019a).

The Every Student Succeeds Act in New Mexico

While both the second and third lawsuits were in progress in New Mexico, in late 2015, as previously discussed, then-President Obama signed ESSA (2015) law. Per the NMPED (n.d.a), New Mexico put itself in good position to continue its prior "successes" under this new policy thanks to its efforts under the NCLB waivers (e.g., meaningful school accountability legislation, a commitment to providing excellent teachers). The state submitted its ESSA plan to the federal government in April 2017, received initial feedback from the Department of Education peer reviewers in June 2017 (USDOE, 2017d), and had the plan fully approved in August 2017 (USDOE, 2017c). In its ESSA plan, New Mexico identified three main academic goals for its students to reach by 2020: 1) 50% (or more) of students to be at grade-level in reading and mathematics, 2) 80% (or more) of students to be graduating from high school, and 3) 75% (or more) of students who graduate from high school and enroll in a higher education institution to not require remediation (NMPED, n.d.a, p. 6; NMPED, 2018a). Additionally, to complement those three academic goals, via Governor Martinez's Executive Order 2016-037, New Mexico adopted an ambitious goal of having 66% of its students earning a college degree (or other postsecondary credentials) by 2030 (State of New Mexico, 2016). As related to accountability, New Mexico's ESSA plan continued the A-F school grade accountability framework, with student growth data to count for the majority of a school's grade, along with school grades being publicly posted online for transparency purposes. The NMPED has also implemented a bonus pay pilot program, where compensation rewards would be based on a teacher's effectiveness rating (NMPED, n.d.a).

Under New Mexico's ESSA plan, the NMTEACH EES remained nearly unchanged from its previous structure, with two main adjustments resulting from stakeholder feedback (see, for example, NMPED, 2016a): student growth data would only count for up to 35% instead of 50% of a teacher's evaluation, and teachers were allowed twice as many absences (i.e., from three to six) before their overall effectiveness rating was negatively affected (NMPED, n.d.a). These changes were solidified in August

2017 when the state’s ESSA plan was officially approved by the U.S. Department of Education, via updated state legislation (NMAC 6.69.8, 2017). In its ESSA plan submission, New Mexico reported utilizing the NMTEACH EES for the past three years had resulted in drastic improvements to its teacher evaluation system and had “rapidly” moved away from too many teachers as being rated as effective (i.e., the widget effect; Weisberg et al., 2009) (NMPED, n.d.a, p. 104; see also Kraft & Gilmour, 2017).

Teacher Evaluation Policy During the Lawsuits

In the following sections, I outline the specific components of the NMTEACH system that was in place from the 2013-2014 to 2016-2017 school years, which also includes the timespan when the three aforementioned lawsuits were filed. The NMTEACH EES that was in effect during those years was the system that was established via NMAC 6.69.8 (2012), and that from which the study for this study were generated (see Chapter 3 for details about the data used for this study, forthcoming).

The guiding principle behind the NMTEACH EES was the NMTEACH Theory of Action (NMPED, 2016b), which “reflect[ed] the belief that if teacher effectiveness improves, then instructional practice will improve, which will then improve student achievement” (p. 4). The NMTEACH EES aimed to measure four broad areas related to teaching: a student’s opportunity to learn, student achievement, a teacher’s instructional quality, and a teacher’s level of professionalism. Each of the main components of the NMTEACH EES (described in more detail, forthcoming) measured one or more of these areas (see Table 1).

Table 1

NMTEACH EES Components and Corresponding Performance Areas

Evaluation Component	Performance Area(s) Measured
VAS Data	Student Achievement
Classroom Observations	Student Opportunity to Learn; Instructional Quality
PPP	Student Opportunity to Learn Instructional Quality Professionalism
Student Perception Surveys	Student Opportunity to Learn
Teacher Attendance	Professionalism

Note: Adapted from NMPED, 2016b, p. 3

A student's opportunity to learn was measured by classroom observations, the PPP component, and the student perception surveys; student achievement was measured by the VAS component; a teacher's instructional quality was measured by classroom observations and the PPP component; and the teacher's level of professionalism was measured by the PPP and teacher attendance components (NMPED, 2016b).

The NMTEACH system was comprised of four main components, with each component contributing to a teacher's overall effectiveness score, though with different weights (see Table 2).

Table 2

Components of Teachers' Evaluation Scores from the 2013-2014 to 2015-2016 School Years

	Teacher in Tested Subjects/Grades		Teacher Not in Tested Subjects/Grades
Years of Student Achievement Data	3+	1-2	0
Evaluation Component			
VAS Data	50%	25%	---
Classroom Observations	25%	40%	50%
PPP	15%	25%	40%
Student Perception Surveys	5%	5%	5%
Teacher Attendance	5%	5%	5%
Total	100%	100%	100%

Note: Adapted from NMPED, 2016b, p. 6

Teachers with three or more years of student achievement data had 50% of their overall evaluation comprised of their VAS scores. The remainder of those teachers' summative evaluation scores was comprised of classroom observations (25%), the PPP component (15%), student or parent surveys (5%), and the teacher's attendance (5%). Teachers with between one to two years of student achievement data had 25% of their overall evaluation comprised of their VAS scores. The remainder of those teachers' overall summative evaluation scores was comprised of classroom observations (40%), the PPP component (15%), student or parent surveys (5%), and the teacher's attendance (5%). For teachers without at least one year of student achievement data, the percentages of the survey and teacher attendance components were unchanged, while the weight of classroom observations increased to 50% and the weight of the PPP component increased to 40% (NMPED, 2016b).

NMAC 6.69.8 (2012) specifically outlined that all aspects of a teacher's evaluation be "based on sound educational principles and contemporary research in effective educational practices" (p. 2). It also specifically stated that the student growth component must be based on "valid and reliable data and indicators of teacher impact on student achievement…" and the observation aspect must incorporate "common research-based observational protocol…that correlates observations to improved student achievement" (p. 3). In the following subsections, I briefly outline the main components of the NMTEACH system in use during the 2013-2014 through 2015-2016 school years.

Student growth data. Student growth data was the driving force behind the NMTEACH system, as the weight of the other evaluation measures were adjusted based on the presence (or absence) of student growth data, as mentioned. The New Mexico VAM was created by Pete Goldschmidt (see Martinez, Schweig, & Goldschmidt, 2016; see also Reiss, 2017), and a teacher's overall VAS score was a weighted average of all possible VAS scores for that teacher from all course groups (i.e., classes taught). That is, for each teacher, a separate VAS score was calculated per subject, per grade, and per assessment (e.g., End of Course [EOC] exam, Partnership for Assessment of Readiness for College and Careers [PARCC] exam [NMPED, 2016b]). The statistical model used to calculate each VAS was purported to control for whether a course had been identified as an "intervention course," the grade level of the student (if a course contains students from multiple grades), and for the proportion of time a student had spent with each specific teacher. Students' testing histories were only included in the model if students had two years of previous testing history (i.e., so expected growth could be calculated). The VAS scores for teachers in Grade 5 and above only included student growth data for students who had data for all of the above datapoints. That is, if a student was missing even one of the aforementioned components, his or her achievement scores were not included in the model used to calculate a teacher's VAS score (see NMPED, 2016b, pp. 14-22 for full model specifics and formulas).

Classroom observations. Classroom observations carried the greatest weight for teachers with fewer than three years of student achievement data, as previously indicated. Teachers were required to be observed at least two or three times per year (depending on the observation plan chosen by each teacher's district), except for teachers who earned at least 73% of all possible total points on their previous summative evaluation (i.e., 146 out of 200 possible points) and at least 50% of all possible total points on their previous evaluation's VAS component (i.e., 35 out of 70 possible points). Teachers who met these criteria were only required to be observed once per year (NMPED, 2016b).

The observation system used in New Mexico is modified from Charlotte Danielson's *Framework for Teaching* (FFT; The Danielson Group, 2013). Of the four domains comprising the FFT, Domains 2 and 3—Creating an Environment for Learning and Teaching for Learning, respectively—were utilized for classroom observations as both focus on the pedagogical aspects of teaching. Both Domains 2 and 3 have five indicators each, combining for a total of 10 indicators. Domain 2 indicators include aspects of teaching like "Establishing a Culture for Learning" and "Managing Classroom Procedures" (NMPED, n.d.c), while Domain 3 indicators include aspects of teaching like "Engaging Student in Learning" and "Demonstrating Flexibility and Responsiveness" (NMPED, n.d.d).

For each indicator, teachers were scored on a five-point scale, with higher values indicating better teaching practices. Generally, a teacher's overall observation score was calculated by taking the sum of the average indicator score (i.e., per indicator) across all observations conducted in a school year. (See NMPED, 2016b, pp. 9-10 for the exact formula used to derive a teacher's summative observation score, which includes adjustments made for missing data). Observations were conducted either by the teacher's principal or assistant principal, or by another certified school personnel member. All observers were required to attend a yearly training to become certified for the classroom observation process (NMPED, 2018b).

Multiple measures. The remainder of a teacher's overall evaluation was comprised of components that were deemed as "multiple measures." These measures consisted of the PPP component, SPSs, and teacher attendance.

Planning, Preparation, and Professionalism (PPP). The PPP component also utilized modified domains from Danielson's (2013) FFT—Domains 1 (Planning and Preparation; NMPED, n.d.b) and Domain 4 (Professionalism; NMPED, n.d.e). As mentioned, the PPP weight also differed by whether or not a teacher had student achievement data (see again Table 2, again). The PPP component of a teacher's overall evaluation was similar to the classroom observation component in that evaluating the PPP component occurred by trained observers, yet these evaluations occurred outside of the classroom (NMPED, 2016b). Domains 1 and 4 had six indicators each, with Domain 1 indicators including aspects of teaching like "Designing Coherent Instruction" and "Setting Instructional Outcomes" (NMPED, n.d.b) and Domain 4 indicators including aspects of teaching like “Communicating with Families” and “Growing and Developing Professionally” (NMPED, n.d.e). The overall summative PPP score was calculated in the same way as the overall summative observation score.

Student perception surveys (SPSs). Regardless of whether a teacher had student growth data, surveys counted for five percent of a teacher’s overall summative teacher evaluation score. Students in Grades 3 through 12 completed their own surveys, while the families of students in Kindergarten through Grade 2 completed the survey in lieu of the students. Each survey contained 10 items that were on a four-point frequency Likert scale, with “0” indicating “Never” and “4” indicating “Always” (NMPED, 2016b). The student survey contained items such as, “My teacher expects me to do my best” and “My teachers checks to see if I understand.” The family survey contained items such as, “My child’s teacher answers my questions” and “My child’s teacher can tell me about my child’s strengths and weaknesses” (NMPED, 2016b, p. 12). The student survey was mapped to the classroom observation rubrics (i.e., Domains 2 and 3; see NMPED, n.d.g), while the family survey was mapped to both the classroom observation and PPP rubrics (i.e., Domains 1 through 4; see NMPED, n.d.f).

The system used to administer and collect survey responses required all 10 items to be answered for a survey attempt to be registered as valid. That is, all survey data used in a teacher’s evaluation was complete per respondent. For a teacher to have received an overall summative SPS score, s/he must have had at least 10 surveys completed by either students or families. Student surveys were given priority in contributing to a teacher’s evaluation, in that if a teacher had both student and family survey data, family surveys were only used if a teacher had fewer than 10 completed student surveys and greater than 10 complete family surveys.

Teacher attendance. Lastly, the teacher attendance metric was calculated in a subtractive method, meaning that a teacher began with the maximum of 10 points and retained all 10 points if s/he was absent for three or fewer days throughout the school year (i.e., the teacher earned 10 out of a possible 10 points). If a teacher was absent for 20 or more days, s/he lost all 10 points (i.e., the teacher earned 0 out of a possible 10 points). For absences between four and 19 days, and for absences across multiple districts, standard formulas were applied to determine the number of points earned (see NMPED, 2016b, p. 14 for formulas used).

Review of the Literature on Teacher Evaluation Measures

As mentioned prior, evaluating teachers to identify their effectiveness has been of chief importance for decades. To date, there have been two main ways to measure teaching effectiveness—VAMs and classroom observations. Combined, these two measures have made up the majority, if not the totality, of many states' and/or districts' teacher evaluation systems over the past 10 years (i.e., since RTTT) (e.g., see Doherty & Jacobs, 2013, 2015). In addition to VAMs and classroom observations, other indicators, such as student perception surveys, SLOs, teacher portfolios, and teacher attendance measures, have begun to gain traction as ways to evaluate teachers as well. In the following sections, I first discuss the history of and documented benefits and concerns about each of the three main measures used to evaluate teachers in New Mexico during the 2013-2014 through 2015-2016 school years: VAMs, classroom observations (of which the PPP component is a part), and SPSs.

A Note About Multiple Measures

Prior to discussing the history of and research on each of the three aforementioned measures, it is worth briefly discussing the push for “multiple measures.” This construal of multiple measures varies from what New Mexico has labeled as “multiple measures,” as outlined above. In New Mexico, “multiple measures” specifically referred to the measures of teacher effectiveness that were not VAMs or classroom observations (i.e., the PPP component, SPSs, and teacher attendance). Here, and regarding general accountability policy, “multiple measures” refers to a variety of different individual components measuring teacher effectiveness (e.g., classroom observations, VAMs, surveys, portfolios) used together within one system (see also Brookhart, 2009), either to collectively inform an overall summative rating of teacher effectiveness or to be combined into one or more composite measures.

The majority of teacher evaluation scholars and researchers, along with a variety of federal policies (e.g., NCLB, NCLB waivers, RTTT, ESSA) and politicians themselves (e.g., Duncan, 2012), have called and continue call for the use of multiple measures to evaluate teachers. Using multiple measures allows for different facets of teaching to be evaluated, as it is generally agreed upon that different measures of teaching effectiveness measure different components that make up the overall “effective teacher” construct. Using multiple measures to evaluate teachers helps to create a more well-rounded and robust picture of a teacher’s effectiveness (Doan et al., 2019; Goe &

Croft, 2009), and also helps alleviate some of the methodological concerns associated with each individual measure (Baker, 2003; Brookhart, 2009; Kane, McCaffrey, Miller & Staiger, 2013; Mihaly, McCaffrey, Staiger, & Lockwood, 2013; Youngs & Grissom, 2016). However, as Doan et al. (2019) note in the case of composite measures, merely combining two or more measures does not automatically result in reduced measurement error. Rather, composite measures are only as reliable and valid as their component measures and can even be less precise than individual measures depending on the reliability and validity of the individual measures (Kane & Case, 2004). As such, caution must still be exercised when interpreting and using multiple measures to evaluate teacher effectiveness.

The use of multiple measures is especially important when teacher evaluations are used for summative purposes (Darling-Hammond, 2012; The New Teacher Project, 2011), as potential methodological concerns are heightened when highly-consequential personnel decisions are at stake (Grissom & Youngs, 2016). Validity of all measures within teacher evaluation systems, and especially convergent-related evidence of validity, has taken on new importance (Sandilos, Sims, Norwalk, & Reddy, 2019). The significance of this importance has been recognized, in part, by the number of research studies that have explicitly examined the relationship among multiple different measures of teacher effectiveness.

However, there still remains many unanswered methodological and pragmatic questions about how to actually implement a multiple measures system that is reliable, valid, and unbiased, as well as fair and transparent for all parties involved. Such questions center around, for example, which measures to include or exclude, what type of weighted model to use, how much weight to give each measure, whether to create one or more composite measures, and the like (see, for example, Brookhart, 2009; Chester, 2003; Martinez et al., 2016; Youngs & Grissom, 2016). While a full in-depth examination into the methodological, pragmatic, and ideological nuances of multiple measures is outside of the scope of this dissertation, the following sections on VAMs, classroom observations, and student perception surveys should be consumed while remembering that each measure is likely to be used within a multiple measures system.

VAM History

As mentioned in Chapter 1, VAMs are used to try to predict or isolate how much "value" a teacher "adds" to (or detracts from) students' achievement. Student achievement is almost always in the form of standardized test scores, and depending on the actual statistical model, VAM developers believe that such models can control for students' prior achievement as well as other potentially confounding student-level demographic (e.g., socioeconomic [SES] status, English language proficiency) and school-level variables (e.g., class size, school capital). Multiple growth models exist today, such as the Education Value-Added Assessment System (EVAAS) (SAS Institute, Inc., 2019), the Student Growth Percentile (SGP) model (see Betebenner, 2009), and other state- or district-specific models, such as the one used in New Mexico (see Martinez et al., 2016; see also Reiss, 2017). Compared to classroom observations (which are discussed in more detail, forthcoming), using VAMs to evaluate teachers is relatively new to the teacher evaluation, especially regarding their massive influence in summative teacher evaluations within the past 10 to 15 years. VAMs, accordingly, have been recently extensively studied by both educational researchers and statisticians/economists, and also hotly debated by policymakers, educators, professional organizations, and the general public (see, for example, AERA Council, 2015; ASA, 2014; Bracey, 2007; Goldhaber & Chaplin, 2015; Pivovarova et al., 2014, Ravitch, 2014). In the following sections, I describe the history and policy importance of VAMs and discuss the current state of research related to student growth measures.

Early history of VAMs. Test-based accountability measures and using assessments for summative purposes in teacher evaluation have been in practice since the 1980s, notably following the release of *A Nation at Risk* (NCEE, 1983), and especially given the increased call for standards-based accountability through standardized testing (Koretz, 1996). During this time, many states began to explore standards and accountability measures, and Tennessee moved to the forefront of the VAM movement. William Sanders, the creator of what is still one of the most popular VAMs on the market, the EVAAS, was an agriculture faculty member at the University of Tennessee teaching advanced statistics courses. He, along with colleague Dr. Robert McLean, had started trying to create assessment methodology incorporating student achievement data, but without some of the noted issues with others' prior attempts (e.g., missing data, different methods of teaching, student movement into and out of classrooms) (Sanders & Horn, 1994). In 1984, McLean and Sanders (1984) published a paper using a model incorporating student achievement data (i.e., standardized test scores) to evaluate teachers. This model would ultimately become the basis for VAMs, and, specifically, Sanders’ Tennessee Value-Added Assessment System (TVAAS). While McLean and Sanders contributed several significant findings relating student test scores and teacher effects, the TVAAS did not gain steam until the early 1990s when the state of Tennessee used the model as a part of its 1992 Education Improvement Act (EIA), which stipulated, in part, that teachers, schools, and districts should be held accountable based on the state’s educational goals (Sanders & Horn, 1994, 1998). Since the TVAAS focused specifically on outcomes, rather than the process by which outcomes were achieved, many saw Sanders’ model as the ideal method to assess accountability per the EIA. Although there was plenty of initial skepticism about the TVAAS in Tennessee, especially regarding its ability to provide “fair, objective, and unbiased estimates” (Sanders & Horn, 1998, p. 248), Tennessee required its use in teacher evaluations statewide, although the state did allow for each individual district to decide how much weight TVAAS results carried within each district’s teacher evaluation system, and cautioned that there needed to be other sources of data included in the teacher evaluation system as well (Sanders & Horn, 1998). At the turn of the century, Sanders retired from the university and worked with SAS Institute, Inc. to provide the TVAAS to any interested state or district that wanted to try to estimate teachers’ effects on students’ learning. He subsequently changed the model’s name to the aforementioned EVAAS.

VAMs in the accountability era. With the passage of NCLB, the focus on student growth data for accountability purposes skyrocketed. Accordingly, as the pressure to improve upon NCLB grew, many believed that VAMs seemed like a possible or even obvious choice for states and districts to tie student performance to teacher evaluation and holding teachers accountable (Doran & Izumi, 2004; Linn, 2004; see also Orland, 2015; Taylor & Tyler, 2012), especially given that teachers should know how much impact they have on a given student's achievement (Braun, 2005). VAMs were also enticing as they were able to connect individual teachers to students, which allowed for the linking and estimating of aggregate teacher effects as per teachers' students' growth in achievement. Since NCLB did not define or specify what "teaching quality" or "effective teaching" meant, it was easy and convenient to associate these constructs with student test scores (Hibler & Snyder, 2015).

Overall, VAMs were initially more prominent at the individual state level rather than across the nation. In 2005, several years after the passage of NCLB, then-Secretary of Education Margaret Spellings announced a new VAM-focused pilot program (USDOE, 2008a). This program allowed states to pilot the use of VAMs in their evaluation and accountability systems as a means to determine whether such models were more accurate for student achievement and accountability purposes compared to AYP calculations. Initially, Tennessee and North Carolina (where SAS Institute, Inc., owner of the EVAAS, was based) were, accordingly, the first two states to participate in the pilot during the 2005-2006 school year, though an additional six states were included in the 2006-2007 school year (USDOE, 2008a). After noted success in the program's first two years, all eligible states (i.e., those that had adequately defined their AYP goals and agreed to share school-level AYP data, among other criteria) were then invited to submit a proposal to participate in the program. This resulted in a total of 15 states participating (USDOE, 2009b).

While some additional states and districts began to incorporate VAMs into their teacher evaluation systems after the aforementioned pilot program, VAMs more commonly became a part of statewide policy efforts with the NCLB waivers in the late 2000s. As previously mentioned, the waivers allowed a state to be excused from certain NCLB mandates provided the state had a plan to improve educational outcomes and the quality of instruction (USDOE, 2012b). To prove they were using rigorous mechanisms for accountability purposes, many states incorporated VAMs into their teacher evaluation systems.

When then-President Obama introduced the RTTT initiative in 2011, VAMs were then cemented as a necessary component of teacher evaluation systems as RTTT explicitly required the use of student growth data to evaluate teachers, provided states wanted to receive the federal funding that was at stake (USDOE, 2009a; Weiss, 2014). Needless to say, and as previously indicated, RTTT had far-reaching effects in terms of "putting growth models on the map" (Collins & Amrein-Beardsley, 2014), so to speak, as the majority of states had not just been using VAMs in their teacher evaluation systems, but had since constructed official policies that actually required the use of VAMs in teacher evaluation systems, along with the requirement that VAMs be used to inform summative decisions about teachers (e.g., teacher termination, awarding or denying merit pay, and the like.

Consequences of the Chetty et al. (2014) study. One of the most highly touted and simultaneously highly criticized research studies that focused on VAMs was published in 2014 (Chetty, Friedman, & Rockoff, 2014a, 2014b), and gained such prominence that it was even cited in the 2012 State of the Union Address (The White House, 2012) when then-President Obama spoke of how an effective teacher can increase the "lifetime income of a classroom by over $250,000" and help children "escape from poverty" (para. 36). Although the potential socioeconomic impact of effective (and ineffective) teachers had already been discussed in the media over the preceding years, the Chetty et al. (2014a, 2014b) study had a profound impact on both policymakers and the general public as many were swayed by the study's findings, which not just touted VAMs, but did so along with all but guaranteeing that good teachers have a profound economic impact on students that affects students across their lifespan. The study was cited in multiple popular media outlets (e.g., CNN [Bennett, 2012], the New York Times [Lowrey, 2012]) thereafter, and scholars and researchers who supported VAMs were also quick to echo their support of the study's findings. The Chetty et al. (2014a, 2014b) study findings were also especially pertinent in several VAM-related lawsuits, such as the notable case of *Vergara v. California* (2012) (Amrein-Beardsley, 2016b) in which the plaintiffs claimed that several of California's statutes violated the state's Constitution, as the statutes led to ineffective teachers being retained and thus negatively and disparately affected minority and low-SES students.

The academic community's support of the Chetty et al. (2014a, 2014b) study was not widespread, as many found the Chetty et al.'s findings to be inaccurate, insignificant, overstated, and/or simply misleading. Noted researchers—some of whom were economists, like Chetty et al.—such as Adler (2013, 2014), Baker (2013), Ballou (2012), Pivovarova et al. (2014), Ravitch (2014), and Rothstein (2015), to name a few, as well as the ASA (2014), came out in staunch opposition to the claims made by Chetty et al. (2014a, 2014b). Adler (2014), for example, mentioned that the judge in the *Vergara v. California* case incorrectly extrapolated and misrepresented Chetty et al.'s (2014a, 2014b) findings, which had major implications in the case as the judged ruled against the state, also due to Chetty's testimony in the case. While the ruling against the state was eventually overturned in an appeals court (see Medina & Rich, 2016), the effect of the use of potentially inaccurate VAM research was highlighted in this case.

Likewise, and just as alarming if not more so, there has been drastic and dramatic reactions to and fallout from the inclusion of VAMs in teacher evaluation systems, especially when data stemming from such models inform the basis of or the entirety of high-stakes consequential personnel decisions for teachers. One of the most potentially damaging issues with VAMs (described in more detail, forthcoming) is their unfair use in consequential personnel decisions. Since 2011, 14 lawsuits related to teacher evaluation and, at times, the use of VAMs, have been filed (Sawchuk, 2015; see also Amrein-Beardsley, 2016a), teachers in large districts (e.g., Chicago) have gone on strike, and the publications of districts' "best" and "worst" teachers have wreaked havoc across the profession (see Gabriel & Lester, 2013b), even driving one teacher to commit suicide (Pathe & Choe, 2013).

ESSA and beyond. With the passage of ESSA (2015) several years ago, states were free to amend their use of VAMs in their teacher evaluation systems, as previously noted. While many states were slow to change their evaluation systems, at least in regard to VAMs, many states have at least since stopped requiring VAMs to be used or have no longer encouraged VAM use (Close et al., 2018). This broad change in VAM use was also likely spurred by the highly publicized arguments in and conclusion of several high-profile lawsuits related to teacher evaluation, such as the *Houston Federation of Teachers v. Houston Independent School District* (2015), *Lederman v. King* (2014), and *State ex rel. Stewart v. New Mexico Public Education Department* (2015) (see Amrein-Beardsley, 2018b). Regardless, at the time of this writing, over 25% of states still use VAMs (i.e., *n* = 14/51; 27.5%; Close et al., 2019)[2]. Until VAMs cease to be used in teacher evaluation systems—regardless of whether their current use is for formative or summative purposes—it is imperative that teachers, school administrators, policymakers, and the greater community understand the purported merits of VAMs, along with the many concerns outlined by researchers and scholars, along with teachers and administrators. Thus, in the following sections, I synthesize the previous research conducted on VAMs as related to teacher evaluation.

VAM Research

The research surrounding VAMs has been incredibly polemic, as the majority of researchers and scholars appear to have very strong views about VAMs (Amrein-Beardsley & Holloway, 2019), both in terms of the statistical and methodological properties of VAMs and VAM use. The literature on VAMs began to proliferate the teacher evaluation research landscape shortly after the introduction of RTTT (USDOE,

[2] In Close et al. (2019), New Mexico is cited as one such state that still uses VAMs. However, per a recent Executive Order (State of New Mexico, 2019) from the state's new Governor, VAMs are no longer a part of the state's teacher evaluation system (see also NMPED, 2019c). Close et al. (2019) cite 15 states as still using VAMs, yet I amended this count to 14 given the cited information from the NMPED.

2009a), with Google Scholar returning over 5,500 results when searching for articles published in 2011 or later with the keywords "[value added models] + [education]."

VAM advantages. One of the biggest asserted benefits of VAMs is that they have the ability to isolate a teacher's actual effects on a student's achievement (Sanders & Horn, 1994; Sanders, Saxton, & Horn, 1997), meaning that VAMs can be used to determine how much of a change (i.e., increase or decrease) in student achievement in a given year can be allotted to an individual student's teacher (versus other external factors). This supposed linking of a teacher's causal inputs to a student's output (in the form of test scores) forms the basis of accountability within teacher evaluation systems, as one of the purposes of teaching is to increase students' academic achievement (see Payay, 2012). Assuming a VAM is indeed able to identify a teacher's individual inputs and link that to a student's outputs, utilizing VAMs allows schools, districts, and states to classify teachers as effective (or ineffective) based solely on their students' achievement (i.e., test scores).

While many who are critical of VAMs cite VAMs' often-biased output (discussed in more detail, forthcoming), VAMs are supposed to function in such a way that they incorporate and therefore control for confounding (i.e., biasing) factors that have been shown to affect student achievement, such as student background characteristics (i.e., demographic data) and prior achievement (Sanders & Horn, 1994, 1998). VAMs' ability to control for such covariates is seen as a strong advantage over other types of models that have been used for accountability purposes (e.g., status or snapshot models, growth models or gain scores; Scherrer, 2011).

From a strict measurement perspective, studies have resulted in researchers concluding that while not perfect, VAMs exhibit high enough levels of reliability (i.e., consistency of estimates over time; described in more detail in the Conceptual Framework section, forthcoming) and validity (i.e., accuracy; described in more detail in the Conceptual Framework section, forthcoming) to warrant them being useful to evaluate teachers (e.g., Glazerman et al., 2010, 2011; Kane et al., 2013; Koedel, Mihaly, & Rockoff, 2015). The majority of the validity-related VAM studies have focused on convergent-related evidence of validity, or "the degree of relationship between the test score and [other] criterion scores" (Messick, 1989, p. 7). Researchers have evidenced this type validity by demonstrating significant and positive relationships between VAMs and other measures of teaching effectiveness, most notably classroom observations (Gallagher, 2004; Harris & Sass, 2009; Kimball et al., 2004; Jacob & Lefgren, 2007; Milanowski, 2004). Other studies in support of VAMs (e.g., Goldhaber & Hansen, 2008; Kane & Staiger, 2008) have focused on countering some of the critical claims against VAMs being unreliable and invalid. For example, Glazerman et al. (2010) conceded that while VAMs were not an error-free measure of teacher effectiveness, the amount of instability and error in VAM estimates were comparable to the error and instability in performance appraisals within other industries.

From a more theoretical perspective, VAMs also are appealing because they are, in theory, based on psychometrically validated assessments (i.e., standardized tests) that allow for more "objective" characterizations of a teacher's level of effectiveness compared to other measures (Doran & Izumi, 2004; Linn, 2004; see also Orland, 2015;

Taylor & Tyler, 2012), such classroom observations, which are quite subjective (discussed in more detail, forthcoming). VAMs also guarantee improvement upon the apparent lack of variation found with classroom observation data (Weisberg et al., 2009), as the nature by which VAM estimates are generated result in a normal distribution of scores (i.e., a bell curve).

Many believe that a normal distribution rather than a negatively skewed distribution is more indicative of the true dispersion of teaching quality (e.g.,Doherty & Jacobs, 2015; Walsh et al., 2017; Weisberg et al., 2009; see also Burgess, 2016, 2017), though largely in part to the rationale espoused by Weisberg et al. (2009). Further, such proponents believe that the normative nature of such a distribution is beneficial for teachers on a practical level as well, as it allows administrators to easily compare teachers (Braun, 2005; Papay, 2012), and teachers themselves can gain "vast new insight[s]" (Glazerman et al., 2010, p. 4) into their own effectiveness, especially as relative to their peers. Lastly, and pragmatically speaking, VAMs are also logistically easy and cost-effective to employ, especially as compared to classroom observations, as student test score data is often readily available due to federal policy requirements, and VAM analyses can be conducted remotely (Rockoff, 2004; Goe, Bell, & Little, 2008).

VAM concerns. The challenges with VAMs are aplenty, as numerous education scholars, researchers, and practitioners have noted. Concerns about VAMs include methodological issues (e.g., reliability, validity, bias), pragmatic issues (e.g., transparency, fairness, appropriate use), and consequential issues (e.g., high-stakes decisions that are informed by VAM estimates) (Amrein-Beardsley, 2014; Amrein-

Beardsley & Holloway, 2019). These measurement and pragmatic aspects are not just limited to VAMs, but rather apply to all educational tests and measurement. That is, per the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), any and all tests, measures, or measurement-related tools in education should adhere to the *Standards*, in that the tests, measures, or tools used are statistically reliable, valid, and unbiased, as well as transparent, fair, and used appropriately.

Such measurement and pragmatic concerns about using student growth data, and VAMs specifically, to evaluate teachers have existed for years. Decades ago, Soar (1973) noted that using student growth data to evaluate teachers seems like "a major oversimplification" (p. 205), especially given the multitude of other factors affecting student achievement (e.g., school effects, peer effects; out-of-school effects such as home environment, health, family income; see, for example, Berliner, 2009). Additionally, using students' test scores for any accountability measure can be incredibly risky as student performance can vary across different tests that are supposed to measure the same constructs or content knowledge, adjusted scores (to account for student background demographics) are not always stable, and tests that have high-stakes outcomes are susceptible to corruption and distortion (e.g., as per Campbell's [1976] Law; Koretz, 1996).

Measurement concerns. A frequently noted problem with VAMs is their (un)reliability, or the noted relative inaccuracy and instability of their estimates. The likelihood of teachers being incorrectly classified (i.e., as "effective" or "ineffective") per their VAM scores, or being classified correctly one year and then reclassified the next, is quite high (Amrein-Beardsley & Geiger, under re-review; Ballou & Springer, 2015; Kersting, Chen, & Stigler, 2012; Koedel & Betts, 2007; McCaffrey et al., 2009; Schochet & Chiang, 2013). Some researchers have noted the likelihood of being misclassified decreases with additional years of data (Corcoran, 2010), but at some point, the improvement in reliability plateaus (Brophy, 1973), regardless of how many additional years of data there are. Overall, studies examining the stability of VAM estimates over time appear to have somewhat mixed or even contradictory findings. Some researchers have indicated that the (corre)linear relationship between VAM scores over time is very weak to moderate[3], at best (e.g., Amrein-Beardsley & Geiger, under re-review; Kane & Staiger, 2012; Koedel & Betts, 2007; Linn & Huag, 2002; McCaffrey et al., 2009; Newton et al., 2010; Sass, 2008), while other researchers have indicated that the relationship can actually be closer to moderate to strong (e.g., Kersting et al., 2013; Loeb & Candelaria, 2012; McCaffrey et al., 2009). However, these stronger relationships appear to be limited to certain grade levels or subject areas (e.g., Goldhaber, Gabele, & Walch, 2012; McCaffrey et al., 2009; Rivkin & Ishii, 2008), which implies that VAM estimates could very well be biased by such factors.

Another grave concern regarding VAMs is whether and to what extent VAMs are valid (i.e., whether they can truly isolate a teacher's effect on a student's level of

[3] I used the following framework to interpret Pearson correlation coefficients: $0.0 \leq r \leq 0.2$ = "very weak;" $0.2 \leq r \leq 0.4$ = "weak;" $0.4 \leq r \leq 0.6$ = "moderate;" $0.6 \leq r \leq 0.8$ = "strong;" $0.8 \leq r \leq 1.0$" = "very strong" (Merrigan & Huston, 2004).

achievement). Proponents of VAMs assume that a student's standardized test score can be directly attributed to his or her teacher (Amrein-Beardsley & Holloway, 2019), yet decades of past research have provided evidence to counter this argument. While teachers might be biggest in-school factor contributing to student achievement, student achievement is also largely affected by a multitude of other in-school and out-of-school factors (e.g., Berliner, 2006, 2009, 2013; Hanushek, Kain, Markman, & Rivkin, 2003; Jargowsky & El Komi, 2011; Lin, 2010; McCoach et al., 2010). Further, it is not possible to isolate exactly how much of a student's achievement gains in a given year can be attributed to a specific teacher versus any other factor (Corcoran, 2010; Goe et al., 2008; Ishii & Rivkin, 2009; Kane & Staiger, 2008; Linn, 2008; Rothstein, 2009). Also worth noting is that if a VAM (or any other measure) is unreliable, it inherently cannot be deemed as valid, as reliability is a requirement for validity (Brennan, 2013; Kane, 2006, 2013; Messick 1975, 1980).

Another largely problematic measurement concern is that VAMs have also been found to be notably biased. In this case, bias is delineated by VAM scores systematically varying for "people belonging to groups differentiated by characteristics not relevant to" the VAM score itself (AERA et al., 2014; discussed in more detail in the Conceptual Framework section, forthcoming). This bias can occur for a multitude of reasons, including how students are sorted (i.e., not randomly placed) into teachers' classrooms and schools (Braun, 2005; Koedel & Betts, 2011; Paufler & Amrein-Beardsley, 2014; Rothstein, 2009, 2010), students' background characteristics (e.g., ELL status, SE status, socioeconomic status; Goldhaber, Quince, & Theobald, 2018; Isenberg, et al., 2016;

Kupermintz, 2003; Newton et al., 2010; Sass, Hannaway, Xu, Figlio, & Feng, 2012; see also Berliner, 2014), students' prior achievement levels (e.g., Rothstein, 2009), the grade or subject level a teacher teachers (Ballou & Springer, 2015; Goldhaber et al., 2012, 2013; Harris & Anderson, 2013; Holloway-Libell, 2015; McCaffrey et al., 2009), class size (Kersting et al., 2013; Sanders et al., 1997), and more.

The issue of bias in VAMs has become so noteworthy that several research studies have been published that illustrate just how severe of a problem this is. One such study was conducted by Rothstein (2010), where he developed what came to be known as the "Rothstein falsification test." Via this process, Rothstein demonstrated that students' fifth grade teachers' VAM scores could be used to predict the same students' achievement in fourth grade, given the bias inherent in the non-random sorting (and tracking) of students into teachers' classes. The association between fifth grade teachers and students' fourth grade achievement was clearly illogical, as future teachers cannot predict students' prior achievement. Rothstein thus concluded that VAMs are highly susceptible to the non-random assignment of students to classrooms (also as discussed above), and also warned of negative consequences if VAM estimates were used for highly consequential personnel decisions. In a second, more recent study, Bitler, Corcoran, Domina, and Penner (2019) demonstrated that teachers add nearly as much "value" to students' height as they do to students' mathematics and reading achievement. While Bitler et al.'s demonstrated relationship between teachers' VAM scores and students' height is due to statistical noise (compared to actual bias; e.g., from the non-random sorting of students), they caution that teacher effects can erroneously be significantly associated with outcomes that "teachers cannot plausibly affect" (p. 3), such as students' height. While Rothstein's (2010) findings were not without criticism (e.g., Goldhaber & Chaplin, 2015; Kinsler, 2012; Koedel & Betts, 2011), both of these studies serve as evidence that strongly call into question the validity of VAMs.

A last troubling concern, which is more related to theory than actual measurement practices (though it does affect measurement practices) is the underlying assumption that the distribution of teacher effectiveness in a given school, district, or state is accurately represented by a normal distribution (Amrein-Beardsley & Holloway, 2019). To generate a teacher's VAM score, teachers are rank-ordered, which results in half of all teachers being rated as below average and half as above average, regardless of how truly (in)effective a given teacher is (Baker et al., 2013; Scherrer, 2011). While teacher quality certainly does vary, both within and between schools, districts, and states, VAM distributions being essentially forced into a bell curve distorts true nature and accuracy of the distribution of teacher quality (Amrein-Beardsley, 2014). This idea, which is discussed in more detail in the Classroom Observations section (forthcoming, this chapter) and in the Implications section (forthcoming, Chapter 6), has the potential to undermine the entire premise of teacher evaluation.

Pragmatic concerns. In terms of fairness, transparency, and appropriate use, VAMs are notably unfair in that only a subset of teachers can be assessed as VAMs are only applicable to teachers who teach in tested subjects (e.g., mathematics, English/language arts) and grades (e.g., middle school). This limitation results in large portions of teachers in a given school, distract, or state not being evaluated by VAMs and thus being evaluated by different measures entirely or different combinations of measures. Prior research has estimated the proportion of teachers who are eligible to be evaluated via VAMs to be as low as 25% and 40% (Gabriel & Lester, 2013a; Harris, 2011; Whitehurst, Chingos, & Lindquist, 2014; see also Amrein-Beardsley, 2014), which results in the majority of teachers being ineligible for VAM-based evaluation. Given VAMs' noted methodological concerns, the discrepancy between evaluation measures across teachers is unfair, especially if VAM estimates are informing highly consequential decisions for some teachers and not for others (Amrein-Beardsley, 2014; Dorans & Cook, 2016; Gill, Bruch, & Booker, 2013; Stecher et al., 2018).

In addition to the issue of fairness, VAM consumers have cited VAMs and the process by which VAM estimates are generated as anything but transparent. Teachers and even administrators have found VAMs, VAM estimates, and even VAM reports that are meant to be consumed by practitioners (versus statisticians) to be incredibly complex and overly complicated, to the point where some have even described VAMs as a "black box" (Derringer, 2010; Gabriel & Lester, 2013a). While many of the companies that own VAMs, such as SAS Institute, Inc. (2019), which owns the popular Education Value-Added Assessment System (EVAAS), claim to provide a plethora of supporting information to help teachers and administrators understand VAMs and interpret VAM estimates (see Sanders, Wright, Rivers, & Leandro, 2009), research has found that teachers and administrators have described VAMs, VAM estimates/outputs, and VAM reports as inaccessible, confusing, not comprehensive, ambiguous, and not formatively helpful (Collins, 2014; Eckert & Dabrowski, 2010; Harris, 2011; Kappler Hewitt, 2015).

In sum, the methodological and pragmatic concerns surrounding VAMs have been so extreme that they have led to multiple lawsuits across the country, as previously mentioned (Amrein-Beardsley & Close, 2019; Sawchuk, 2015; Pullin, 2013). The foci of these lawsuits have been widespread, in that a variety of aspects of VAMs have been contested. For example, VAM-related lawsuits have covered personnel decisions (e.g., termination; *Houston Federation of Teachers v. Houston Independent School District*, 2015), the publication of teachers' evaluation scores and rankings (*Mulgrew v. Board of Education of the City School District of the City of New York*, 2011; Song, 2012), and the relative weight of VAM scores compared to other measures in teacher evaluation systems (*New York State United Teachers Association v. Board of Regents of the University of the State of New York*, 2011), among others. Although VAMs are no longer a requirement in teacher evaluation systems, as previously mentioned, they remain a hotly contested measure of teacher effectiveness that are still being utilized in some states and districts. While some, notably economists (e.g., Chetty et al., 2014a, 2014b; Amrein-Beardsley & Holloway, 2019), have tried to make cases for VAM use, the majority of research surrounding VAMs seems to be more concerned about their use in teacher evaluation systems than in favor.

Classroom Observation History

Classroom observations have been used for decades to evaluate teachers. Classroom observations entail a principal or other trained or certified professional going into a classroom to observe a teacher teaching, while often comparing the teacher's behaviors and actions against a predefined standards-based rubric. Classroom observations as a measure for teacher effectiveness is the most common evaluation method (Dorety & Jacobs, 2015; Steinberg & Donaldson, 2016), and research has found that observations have many benefits, including being able to measure facets of teaching that are difficult to assess in any other manner (Soar, 1973). In the following sections, I describe the history and policy importance and discuss the current state of research related to classroom observations.

Early history of classroom observations. Early conceptualizations of what we now consider to be classroom observations of teachers began approximately two hundred years ago, in the 1800s, when school administrators or principals would observe teachers' instruction. However, recognizing a link between teachers' behaviors and potential student outcomes did not truly emerge until the early 1900s with the work of Taylor, Thorndike, and Cubberley, who worked under the premise that certain behaviors (i.e., inputs) could produce more efficient or better student outcomes (i.e., outputs) (Marzano, Frontier, & Livingston, 2011). Even after this input-output link was recognized, the majority of classroom observations in the first half of the 20th century were more unofficial, formative, and low-stakes than formalized summative high-stakes procedures that were legislated via state, district, or school policy (Cohen & Goldhaber, 2016; Hibler & Snyder, 2015).

Between the 1940s and 1970s, the literature on teachers and teaching began to identify teachers as individuals with unique skills who could directly affect individual student outcomes (Marzano et al., 2011) and some researchers started to conceptualize how to best evaluate teachers using more objective and systematic frameworks (Brophy

& Good, 1986; Ellett & Teddlie, 2003), although Soar (1973) and others noted that teaching and its effects were complex phenomena and therefore difficult to accurately and adequately measure. At this time, regardless, the majority of teacher evaluations were based on classroom observations, though educational and other researchers were trying to explore a more causal link between effective teaching and student outcomes.

The shift to standardization. In the last quarter of the 20th century, classroom observation protocols began to gradually shift to be more objective and standardized in nature, as some evaluation systems started to incorporate checklists into the observation frameworks, in addition to using more standardized measures to license teachers (Ellett & Teddlie, 2003). During this time, an influx of clinical supervision models permeated the education sphere, which were akin to models used in teaching hospitals where there was a relationship between the supervisor and teacher that was guided by observations and discussions to improve the growth and effectiveness of both parties (Goldhammer, 1969). The five phases of Goldhammer's model—pre-observation conference, classroom observation, analysis, supervision conference, analysis of the analysis—set the stage for many subsequent classroom observation frameworks.

In the 1980s, amid the "manufactured crisis" (Berliner & Biddle, 1995) stemming from the infamous *A Nation at Risk* report (NCEE, 1983), teacher evaluation came into the spotlight as the NCEE report highlighted teachers as those who had the most influence on students and specifically, student's achievement. Around the same time, there was a push to further professionalize the teaching profession (Darling-Hammond & Schlan, 1992), especially as related to teacher certification, and hundreds of policies were enacted and implemented to this end (Darling-Hammond & Berry, 1988). Further, Wise et al. (1984) released a report regarding teacher evaluation and noted that the then-current evaluation practices were often lacking when it came to the information needed for teachers to actually improve their practices. Wise et al.'s report featured several recommendations for improving teacher evaluation systems, which most notably included improved training for evaluators, increased time for evaluations, and standard competencies under which teachers would be assessed. Up until that time, however, instruments used for the purposes of observation were often not grounded in any sort of teaching-related theory or research (Porter, Youngs, & Odden, 2001).

Standards-based observations began to gain traction in the 1990s (Milanowski, 2004), which was a testament to the political landscape that was focused more on accountability and reform—especially as related to teacher quality and evaluation (Ellett & Teddlie, 2003). By the 1990s, the majority of all states and D.C. had formal teacher evaluation policies, and the majority of those states/D.C. required teachers to be observed within their classrooms on a regular basis (Valentine, 1990 in Darling-Hammond & Schlan, 1992).

In 1996, Charlotte Danielson released the first edition of her framework, the FFT (The Danielson Group, 2013), which is now one of the most common and frequently used observational frameworks in the country (Pianta & Hamre, 2009). Danielson was one of the first scholars who tried to accurately capture the complexity and idiosyncrasies of teaching within her framework, specifically noting that the model incorporated what she conceptualized and continued to assert as the multiple stages of teaching (Danielson, 1996; Marzano et al., 2011). At the turn of the 21st century, observations were still in widespread use across the country, though there was scant evidence to indicate classroom observations resulted in tangible consequences, such as improving instruction (Peterson, 2004; Stronge & Tucker, 2003). Further, at this time, VAMs were beginning to be incorporated into teacher evaluation systems, which drastically changed the teacher evaluation landscape, as described previously.

Observations in the accountability era. From a policy standpoint, the introduction of NCLB in the early 2000s and RTTT in the early 2010s allowed complex standards- and research-driven observational frameworks, such as the FFT (The Danielson Group, 2013), the Marzano Teacher Evaluation Model (Learning Sciences International, 2017), and The System for Teacher and Student Advancement, or TAP (National Institute for Excellence in Teaching [NIET], 2017), to gain more steam, as policies and initiatives pushed for more data-driven evaluation of teachers and holding teachers accountable for their performance (Kane, Taylor, Tyler, & Wooten, 2011; Hamre et al., 2013). While the majority of the accountability focus was realized via VAMs, classroom observations remained an integral component and consistent feature of evaluation systems. This remained so until the release of *The Widget Effect* report (Weisberg et al., 2009) report, which, as mentioned prior, severely critiqued the accuracy and usefulness of classroom observations, especially from an accountability perspective.

***The Widget Effect* and subsequent policy implications.** When *The Widget Effect* (Weisberg et al., 2009) was released, the use of standards-driven observational systems has been supported by federal policies and initiatives like NCLB and RTTT, both of which called for an increase data-driven protocols to improve teaching (Weber, Waxman, Brown & Kelly, 2016), as mentioned prior. While research on multiple aspects of classroom observations (e.g., proper use, measurement properties, the relationship among observations and other teacher effectiveness measures) had been conducted for decades, *The Widget Effect* (Weisberg et al., 2009) had possibly the most profound effect on teacher evaluation policy at the time, as well as on many education stakeholders' and the public's opinion of teachers and teacher evaluation systems.

In the report, Weisberg et al. highlighted the lack of variation in teachers' evaluation scores and this lack of variation—termed the "widget effect" (p. 4)—was such that the majority of teachers were identified as either "effective" or "highly effective" with an incredibly small percentage of teachers being deemed "ineffective." Weisberg et al. claimed this lack of apparent variation had potentially dire consequences, in that teachers who truly were ineffective were being terminated "with exceptional infrequency," which was a part of the "fundamental crisis" in education (p. 2). To reach this conclusion, Weisberg et al. studied nearly 15,000 teachers and 1,300 administrators across 12 districts in four states, and deduced that, in evaluation systems where non-binary categories were used, 94% of teachers were classified as "effective" or "highly effective." In systems where binary category were used, the percentage of "effective" or "satisfactory" teachers jumped to 99%. Weisberg et al. noted that this lack of variation had several consequences, including failing to recognize teachers who truly were excellent at their jobs; failing to provide adequate professional development or formative feedback to teachers who needed it the most; failing to address issues of poor or ineffective teaching; and missing the opportunity to provide extra support to beginner teachers, who have been shown to need such support the most.

Some researchers (e.g., Chetty et al., 2014a, 2014b; Hanushek, 2011; Kane, 2015) quickly jumped on board with Weisberg et al.'s (2009) findings by agreeing that teacher evaluation systems, as then structured, did not do an adequate job of identifying truly effective or ineffective teachers, and therefore did not reward or punish teachers appropriately (Taylor & Tyler, 2012). Further, Weisberg et al.'s (2009) findings also seemed to confirm what some principals already believed: the distribution of teacher quality should be closer to normal rather than highly skewed (Jacob & Lefgren, 2007). The fallout from *The Widget Effect* and subsequent policy implications were far-reaching. While the report increased the perceived need for VAMs, it also helped shape the belief that there should be a positive (and ideally statistically significant) relationship among the individual teacher evaluation measures within one teacher evaluation system. The push for both a more normal distribution of teacher effectiveness and alignment between scores has been so strong that one observational system—the System for Teacher and Student Advancement (i.e., TAP; NIET, 2017)—has used its purported ability to transform teacher quality distributions from negatively skewed to more normal as a marketing tactic (Jerald & Van Hook, 2011).

The belief in the "alignment" of different teacher effectiveness measures—with VAM scores being the measure upon which other measures should be aligned—has also resulted in multiple instances of observation data being overtly or covertly manipulated so observation scores fell more in line with teachers' VAM scores (i.e., artificial conflation; Amrein-Beardsley & Geiger, 2019b; Collins, 2014; Poon & Schwartz, 2016). Further, some states (e.g., Alabama, Georgia, Tennessee) tried to pass legislation to require such alignment, and Tennessee succeeded in actually requiring the artificial convergence of VAM and observational scores via state policy (Tennessee State Board of Education, 2012; see also Amrein-Beardsley & Geiger, 2019b). In addition to the push for alignment across states, several districts (e.g., Baltimore County School District, Houston Independent School District [HISD]) also attempted to strong-arm their administrators into providing aligned scores, with principals feeling pressured to ensure that their observational ratings of teachers matched the teachers' VAM ratings (Collins, 2014; Goldring et al., 2015; HISD, 2012, 2013).

Notwithstanding the critiques of classroom observations (which are described in more detail, forthcoming), and especially the concern about the apparent misalignment between VAM scores and observation scores, classroom observations remained a stronghold in teacher evaluation systems. Further, they were measure that largely informed teachers' evaluations from a summative standpoint (Steinberg & Donaldson, 2016), and they currently remain in use across the country today. In the following sections, I outline some of the benefits of and concerns around using classroom observations to evaluate teachers.

Classroom Observation Research

Classroom observations allow teachers' actual behaviors and practices to be observed and described within their daily natural settings (Hamre et al., 2013; Ross, Smith, Alberg, & Lowther, 2004), which is one reason they are so popular with teacher evaluation systems. Observations are also are able to measure a variety of aspects of teaching, such as a teacher’s engagement levels, student-teacher interactions, and other nuanced effective teaching practices that other measures, like VAMs, cannot (Kane et al., 2011; Soar, 1973; Weber et al., 2016). While classroom observations and related observational frameworks have not undergone as intense scrutiny as VAMs, there still has been substantial research in the recent past assessing observations’ benefits and concerns, though some note (e.g., Polikoff, 2015) that there is still much more room for further inquiry into the various methodological and pragmatic aspects of classroom observations.

Observation advantages. The benefits to classrooms observations are many, and as such they are widely accepted as a valuable measure within teacher evaluation systems as both administrators and teachers find them familiar, transparent, and straightforward (Garrett & Steinberg, 2015). Generally speaking, observations are a useful means of formative feedback for teachers, who can then subsequently improve their instruction; helpful for evaluating facets of teachers which other measures are unable; and are seen as relatively robust from a measurement perspective, especially when compared to other measures, such as VAMs.

Compared to other measures that were not in existence prior to the late 1900s (such as VAMs), observations were seen as easy and practical to implement. While observations and the frameworks in which they are a part are currently often more complex and standards-driven in nature than when they were initially developed (Cohen & Goldhaber, 2016), observations are viewed by many (e.g., administrators, policymakers, and teachers themselves) as a highly accepted way to measure teaching effectiveness (Goe et al., 2008), with little to no pushback from said stakeholders, or from the academic or research communities.

One of the most tangible benefits of classroom observations their ability to provide teachers with formative feedback that can then be used to improve teaching practices (Hill & Grossman, 2013). This is in direct contrast to other measures, such as VAMs, which teachers have found to be unhelpful from a formative sense (Collins, 2014; Eckert & Dabrowski, 2010; Harris, 2011; Kappler Hewitt, 2015). Further, the direct and individualized feedback for teachers from which teachers can adjust their instruction as needed to improve their teaching techniques (Kane & Staiger, 2012; Weber et al., 2016) ultimately benefits students (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Ross et al., 2004; Taylor & Tyler, 2012). This benefit is also potentially realized years later, as Taylor and Tyler (2012) found evidence to support the idea that teachers who experienced classroom observations had students show greater achievement gains the year of said observations, along with several years after the observations as well. These overall benefits from observations are best obtained when classroom observational systems are standards-based and the actual in-class observations occur multiple times throughout a school year (Steinberg & Sartain, 2015; Taylor & Tyler, 2012).

Additionally, one of the most far-reaching benefits of classroom observations is that all teachers, regardless of subject or grade taught, can be assessed. Unlike VAMs, which can only be used for a small proportion of teachers, classroom observations are frequently the main component of teacher evaluation systems for teachers who teach non-tested subjects and/or grades (Garrett & Steinberg, 2015; Weber et al., 2016). As such, overall, classroom observations are seen as both a clear and fair way to evaluate all teachers (Stecher et al., 2018).

From a measurement perspective, observations have been found to be both relatively reliable and valid. Observations are thought to be quite stable (Cohen & Goldhaber, 2016), though several researchers have noted that more research needs to be conducted specifically regarding the stability of observations (Polikoff, 2015; Whitehurst et al., 2014). Of the studies that have been conducted regarding the reliability of observational measures, overall, observations have been found to be a reliable way to measure teaching practices, provided there are multiple observations per year (Hill, Charalambos, & Kraft, 2012; Ho & Kane, 2013; Sartain, Stoelinga, & Brown, 2011; Smolkowski & Gunn, 2012), and especially when compared to value-added scores (Polikoff, 2015; Whitehurst et al., 2014).

Classroom observations have also been found to be valid, both in terms of face validity and convergent-related evidence of validity (see Messick, 1989). This face validity is likely a strong factor behind teachers and administrators accepting classroom observations as a useful measure to evaluate teachers, as teachers and administrators believe that observations are measuring what they are intended to (Cohen & Goldhaber, 2016). Regarding convergent-related evidence of validity, many studies in the recent past have evaluated the relationship between standards-based observational data and other measures of teaching effectiveness. Overall, researchers have determined that observational scores have a strong relationship with teachers’ effectiveness ratings as predicted by VAMs (Grossman et al., 2010, 2014; Harris, 2011; Hill et al., 2011; Kane & Staiger, 2012; see also Polikoff & Porter, 2014; Rothstein & Mathis, 2013), as well as with student achievement (Kane et al., 2011). However, it should be noted that this determination is viewed by some as both potentially inaccurate and subjective. For example, Garett and Steinberg (2015) noted that convergent-related evidence of validity cannot truly be accurately established due to the non-random sorting of students in classrooms, while Amrein-Beardsley (2014) has noted that the correlations between teachers' VAM and observation scores have often only been moderate, at best, thus calling into question others labeling the relationship between VAM and observation scores as "strong."

Observation concerns. Although classroom observations have been in use for decades and are not nearly as contentious as VAMs (Cohen & Goldhaber, 2016), like any assessment tool, they have their share of methodological and pragmatic concerns. Hazi and Rucinski (2009) succinctly labeled traditional methods of teacher evaluation, such as observations, as "flawed, contested, and problematic" (p. 3), especially in the era of accountability. The main challenges with observations revolve around measurement issues, such as reliability and bias, although pragmatic concerns, such as lack of efficiency and high cost (Pianta & Hamre, 2009), have been noted as well.

***Measurement concerns*.** One of the most prominent measurement concerns with observing teachers in a classroom is the subjectivity of observations (Papay, 2012). While observational systems have generally been accepted as a valid way to measure teaching effectiveness, the validity rests on not just the framework guiding the observations, but also on those conducting the observations (Noe, Tocci, Holtzman, & Williams, 2013). Without proper training and calibration of the observers, observational systems are not reliable or valid (Johnson, Penny, & Gordon, 2009; Noe et al., 2013). Even when standardized frameworks, protocols, or rubrics are used to guide the evaluations of teachers, the subjectivity of the observers can still lead to high error rates. Prior research has found error from the observers themselves to account for anywhere from 25% to as much as 70% of the variance in observational scores, depending on the observation framework used (Casabianca et al., 2013; Curby et al., 2011; Hill et al., 2012; Ho & Kane, 2013; see also Mashburn, Downer, Rivers, Brackett, & Martinez, 2014).

In a study of the stability of observational scores conducted over 40 years ago, Brophy, Coulter, Crawford, Evertson, and King (1975) found substantial intertemporal variability for certain sub-scales used in observational evaluations. More recently, within some observation protocols, some scales of teaching constructs were found to be quite stable over time yet while others fluctuated dramatically (Garrett & Steinberg, 2015; Polikoff, 2015). This instability and variation over time has also been problematic in terms of how teachers are classified based on their observation scores. For example, when using criterion-referenced scores (i.e., cut scores) to determine different levels of teaching effectiveness, which many observational frameworks employ, there can be a large likelihood of teachers receiving a certain classification one year and then a different classification the following year (Hiebert & Morris, 2012; Polikoff, 2015). Like with VAMs, these changes in effectiveness classifications can be incredibly concerning, especially in states or districts where observation scores heavily inform summative decisions for teachers (Polikoff, 2015).

Numerous types of rater biases also contribute to the subjectivity of classroom observations (Hoyt, 2000; Wherry & Bartlett, 1982), most of which stem from the natural tendency for people to hold implicit or subconscious stereotypes or biases about others (see, for example, Greenwald & Banaji, 1995). Regarding classroom observations, an observer can have either overt or subconscious thoughts or ideas about a given teacher (see Goe et al., 2008), which can, generally speaking, lead the observer to fall victim to the halo effect (i.e., observers rate a teacher highly on a particular item or construct due to a perceived overall good impression about that teacher; see Thorndike, 1920) or the fatal-flaw effect (i.e., the opposite of the halo effect; observers rate teachers poorly due to a perceived overall negative impression about the teacher) (Noe et al., 2013). The halo and fatal flaw effects can also be compounded by observers being biased due to the background characteristics of the students in a teacher's classroom (e.g., Mason, Gunersel, & Ney, 2014; McGrady & Reynolds, 2013), which can lead to even further biased observation scores for a given teacher.

Other types of biases that can affect teachers' observation scores include the familiarity bias (i.e., observers unknowingly basing some part of their rating on previous interactions or experiences with a given teacher), drift (i.e., the tendency of observers' scores to shift, overall, over time), and the central tendency effect (i.e., observers are reluctant, for a possible multitude of reasons, to give ratings that are at the extremes of a scoring scale) (Noe et al., 2013). Other factors that can bias a teacher's observation score include poor inter-rater reliability (if teachers are observed by more than one evaluator); whether an observer views the concept of classroom observations and teachers' practices more from an accountability lens or from a developmental framework (Bell, Jones, Qi, & Lewis, 2018; Gabriel, 2018; Gabriel & Woulfin, 2017); observers only rating teachers whose teaching practices are familiar to them; observers rating subject or grade teachers in which the observers themselves are not familiar with the subject content or grade standards; the background characteristics of the students within a teacher's classroom (e.g., student demographics, prior achievement; e.g., Blazar, Litke, & Barmore, 2016; Borman & Kimball, 2005; Campbell & Ronfeldt, 2018; Chaplin, Gill, Thompkins, & Miller, 2014; Gill, Shoji, Coen, & Place, 2016; Steinberg & Garrett, 2016; Whitehurst et al., 2014); and background characteristics of the teachers as related to the observers (e.g., gender, race/ethnicity; e.g., Bell et al., 2012; Whitehurst et al., 2014). In essence, almost all aspects of the observation of classroom teachers—including observer, teacher, and student demographics and the observation frameworks, rubrics, or protocols—can result in unreliable and biased output.

Lastly, and worth noting given its potentially dire consequences, is the purposeful manipulation of teachers' observation scores (i.e., artificial conflation; Amrein-Beardsley & Geiger, 2019b). As previously discussed, the driver behind this manipulation is the idea that the true distribution of teaching quality is a bell curve, and too many teachers are rated too highly (Weisberg et al., 2009). This manipulation, which is possibly one of the most pernicious actions regarding the adjustment of scores, can occur either by observers in real time or by principals after the fact, such as to align teachers' observation scores with their VAM scores. Such manipulation is so easily possible given the subjective nature of classroom observations. Not only does this manipulation do an obvious disservice to teachers, but it also results in a teacher effectiveness measure (and overall system) that is invalid at best and rife with a variety of measurement errors, along with dangerous accountability and policy implications, at worst.

Pragmatic concerns. From a practical perspective, the most frequent concerns about observations are that they are time consuming to conduct (Larkin & Oluwole, 2014), especially on a large scale, and not cost effective or as affordable as other measures (Rothstein & Mathis, 2013). The number of teachers who need to be evaluated in a given term or year compounded with the time and cost needed to evaluate each teacher can easily result in observations that are too brief to adequately determine a teacher's effectiveness or provide each teacher with truly useful feedback (Goldrick, 2002). Further, in some states or districts, veteran teachers and/or teachers with tenure are not even observed on a yearly basis (Peterson, 2004; Weisberg et al., 2009). For example, prior to the initial NMTEACH EES system being implemented in 2012, New Mexico only required its veteran teachers to be observed once every three years (USDOE, 2011c).

From a pragmatic perspective, infrequent observations can result in stagnant teacher practices, which can subsequently result in a lack of instructional improvement (Hill & Grossman, 2013); ineffective teachers being allowed to remain in the profession; and/or possible due process concerns if observations are used for summative purposes (Larkin & Oluwole, 2014). From a measurement perspective, infrequent observations negatively affect the reliability of observation measures along with observational systems as a whole (Kane & Staiger, 2012). Overall, classroom observations remain a powerful tool within teacher evaluation systems, although there are some notable measurement and pragmatic concerns, especially surrounding the artificial manipulation of observation scores, which should not be taken lightly.

Student Perception Survey (SPS) History

Student perception surveys (SPSs) are used to obtain students' opinions about their teachers' in-classroom teaching practices and socioemotional qualities as related to instruction. SPSs are useful for either formative purposes (e.g., to guide professional development) or to inform part of a teacher's overall summative effectiveness rating (Schulz, Sud, & Crowe, 2014). While surveys to evaluate teachers for formative purposes have been used for decades, this has primarily occurred in higher education and only sporadically in elementary and secondary education in a handful of states or districts (Marsh, 1987, 1991, 2007; Marsh, Dicke, & Pfeiffer, 2019; Peterson, Wahlquist, & Bone, 2000). However, as of late, and especially given the push for multiple measures, as indicated previously, there has been a recent increase in SPS usage in teacher evaluation systems as SPSs offer several unique benefits that no other (current) measure provides. In the following sections, I describe the history and discuss the current state of research related to SPSs.

Early history of SPSs. When discussions surrounding teacher evaluation began to take off during the MCT era, researchers had been lobbying for the use of multiple evaluation measures other than just classroom observations or student outcomes (e.g., academic achievement, graduation rates). Noted researchers, such as Glass (1974) and McGreal (1983), cited student evaluations as recommended inputs in the evaluation of teachers. In the 10 years between the NCEE releasing its *A Nation at Risk* report in 1983 and the early 1990s, it was estimated that SPSs were being used in under five percent of districts nationwide (Educational Research Service, 1988; Loup, Garland, Ellett, & Rugutt, 1996). Although SPS usage was sparse at the time, researchers (e.g., Aleamoni, 1999; Peterson, Stevens, & Ponzio, 1997; Peterson et al., 2000; Stronge & Ostrander, 1997) continued to examine SPSs.

The first SPS. In the early 2000s, Harvard economist Ronald Ferguson ended up creating what would eventually become the first commercially available SPS in the country, the Tripod student survey (Tripod Education Partners, 2017; Wallace, Kelcey, & Ruzek, 2016). Ferguson was helping a small school district in Ohio with an unexplained issue of "uneven student achievement" (LaFee, 2014, p. 18). Ferguson ended up anonymously surveying the districts' students about their experiences with their teachers when none of the typical methods of teacher evaluation led to any conclusions about the odd achievement data. Ferguson was pleasantly surprised by the survey results as he found that the students were able to recognize effective and ineffective teaching, their answers were accurate and consistent, and they took the survey seriously (LaFee, 2014). From there, Ferguson (2012) worked with the Ohio district for several years to further refine this first SPS, which he later named the Tripod, and between 2001 and 2012, nearly one million students had taken the survey.

SPSs in the accountability era. The use of SPSs within contemporary teacher evaluation systems increased again after *The Widget Effect* (Weisberg et al., 2009) report was released and highly publicized, along with the implementation of the RTTT initiative (USDOE, 2009a). The combination of Weisberg et al.'s (2009) findings, a new multi-billion dollar grant competition, and the general discourse and rhetoric surrounding teacher accountability at the time led states to consider measures to evaluate teachers other than VAMs (which were already mandated per RTTT) and the more traditional classroom observations (Schulz et al., 2014).

The Measures of Effective Teaching study. While SPS usage had increased in the first decade of the 21th century compared to decades prior, and SPSs were increasingly seen as valuable measures to be included in a teacher evaluation system, what really spurred more widespread use was the Measures of Effective Teaching (MET) study. The MET study ran from 2009 to 2011, and its stated purpose was to identify measures of effective teaching that could inform administrators, schools, districts, and states about teachers' strengths and weaknesses (Bill & Melinda Gates Foundation, n.d.) through assessing the reliability and validity of said measures. Its purpose also included determining if such measures could be utilized together in one evaluation system that would be "fair, valid, and reliable" (White & Rowan, 2014, p. 5). The main foci of the MET study were VAMs and classroom observations, as they were the two measures most commonly used to evaluate teachers at the time, but study researchers also used data from Ferguson's Tripod SPS to assess its reliability and, specifically, convergent-related evidence of validity (Raudenbush & Jean, 2014). Study results found the Tripod to be a reliable and valid way to measure teacher effectiveness, and researchers also indicated that SPS data were positively linked to teachers' VAM scores (Bill & Melinda Gates Foundation, 2012). While the MET study drew its share of criticism (see, for example, Jensen et al., 2019), its results were highly publicized, which helped to popularize SPSs.

Current state of SPS usage. Shortly after the MET study findings were published (see Kane, Kerr, & Pianta, 2014) and highly publicized in national and local news sources (e.g., Butrymowicz, 2012; Heitin, 2012; O'Donnell, 2014), more states and districts began to incorporate SPSs within their teacher evaluation systems and also use them for summative purposes. At this time, the NCTQ also began tracking state policy regarding the use of SPSs (Doherty & Jacobs, 2015). In 2013, nearly one quarter of all states (i.e., *n* = 12/50) either required or allowed SPSs. In 2015, the percentage of states requiring or allowing SPSs had increased to over 65 percent (i.e., *n* = 33/50), and by 2017, 34 states (i.e., 68%) required or allowed SPSs (Ross et al., 2017). However, this number slightly decreased in 2019, with only 31 states requiring or allowing SPSs (Ross & Walsh, 2019).

Since the development of the Tripod in the early 2000s and the subsequent MET study a decade later, several other companies have developed commercially available SPSs, and many states and districts have created surveys specifically for their own uses. However, broadly speaking, little is still known about SPSs (Geiger & Amrein-Beardsley, 2019; Marsh et al., 2019) in terms of their potential benefits, methodological issues, best uses within teacher evaluation systems, function within the greater accountability landscape, and the like. In the following sections, I synthesize the research that has been conducted on SPSs as related to teacher evaluation and outline SPSs' cited benefits and concerns.

Student Perception Survey (SPS) Research

To date, the majority of empirical research about SPSs within contemporary teacher evaluation systems uses the survey data from the MET study (i.e., the Tripod survey data; see Kane et al. 2014). Data from surveys other than the Tripod have either not been empirically analyzed at all, have not been empirically analyzed in conjunction with other measures of teaching effectiveness, and/or have not been empirically analyzed with the same frequency as the Tripod data. This is likely due to the fact that the Tripod data, along with the entirety of the MET data (e.g., teachers' VAM scores, classroom observation scores), is available to academics for research purposes (Inter-university Consortium for Political and Social Research [ICPSR], 2019), so it is easier to access and therefore examine compared to data from other SPSs. It is important to keep this point in mind when consuming SPS-related research, as one should be careful about making and/or believing sweeping generalizations about SPSs if SPS research has, to date, been from a single survey tool.

SPS advantages. Although SPSs are relatively new to contemporary teacher evaluation systems, proponents of SPSs see many benefits in their use. From both a measurement and pragmatic standpoint, one of the biggest benefits of SPSs is that they target the very students with whom teachers interact the most and who are the intended targets of teachers' instruction (Bill & Melinda Gates Foundation, 2012; Ferguson, 2012; Follman, 1992; Goe et al., 2008; LaFee, 2014). SPSs allow students' voices to be heard, which is of high value as students are the main stakeholders in their own educations (Wiggins, 2011). Further, student feedback provides teachers with a different perspective than, for example, what administrators provide via classroom observations (Balch, 2016), and teachers see this different perspective as relevant to teacher effectiveness evaluation (Stecher et al., 2018).

From a measurement perspective, some research—much of it resulting from the MET study and/or Tripod data—has indicated that SPSs are a reliable and valid measure of teacher effectiveness, overall and especially in relation to other measures (e.g., Balch, 2016; Kuhfeld, 2017; Rowley, Phillips, & Ferguson, 2019; Wallace et al., 2016; van der Scheer, Bijlsma, & Glas, 2019). For example, the developers of the most popular SPS (i.e., the Tripod) and classroom observational system (i.e., the FFT), respectively, provided evidence that both measures assess similar constructs relating to teacher effectiveness (Ferguson & Danielson, 2014). SPSs have also been found to be indicative of predictive validity regarding student achievement gains (Kuhfeld, 2017; Wilkerson, Manatt, Rogers, & Maughan, 2000), which might imply that SPSs might also be predictive of outcomes in subjects and grades that are untested (Raudenbush & Jean, 2014), and teachers' overall level of teaching effectiveness (Worrell & Kuterbach, 2001). From a theoretical perspective, SPSs can also be more reliable than other measures of teacher effectiveness, such as classroom observations, as students spend multiple hours with their teachers, compared to classroom observers who might spend only a few hours with each teacher, at most. In addition, unlike classroom observations, which are typically conducted by only one or two observations, SPSs are averaged across many students, which is likely to result in better reliability and validity (Geiger & Amrein-Beardsley, 2019). Several additional studies have also evidenced similar findings related to SPS reliability (Fauth, Decristan, Rieser, Klieme, & Büttner, 2014; Ferguson, 2008; Follman, 1995; Peterson et al., 2000; Wagner, Gollner, Helmke, Trautwein, & Ludtke, 2013; Wallace et al., 2016) and validity (Kane & Cantrell, 2010; Kane & Staiger, 2012; Waxman & Eash, 1983), though not all of these studies occurred within the U.S.

Other, non-measurement benefits to SPSs are that they are relatively cost-effective (Bill & Melinda Gates Foundation, 2012; Kyriakides, 2005; Schulz et al., 2014), especially as compared to the cost of building the data systems required for VAMs use and analysis and classroom observations, respectively. Surveys are also easy to administer and provide states, districts, and/or schools a mechanism by which they can gather a quick impression of students' opinions (Schulz et al., 2014). Additionally, SPSs can provide feedback to teachers that is both prompt and formative, which other measures, such as VAMs, are unable to. This prompt formative feedback can result in teachers updating their pedagogical practices and socioemotional behaviors before the school year is over, which ultimately results in better academic achievement and classroom experiences for students (Bill & Melinda Gates Foundation, 2012; Stevens, Harris, Liu, & Aguirre-Munoz, 2013; Whitehurst et al., 2014).

SPS concerns. Although the research on SPSs at the elementary and secondary level has been relatively sparse, researchers have documented concerns about using SPSs to evaluate teachers. Like VAMs and classroom observations, potential issues with SPSs include those from both a measurement and pragmatic perspective. Overall, while SPS usage across the country is relatively widespread at the moment (see Ross & Walsh, 2019), researchers agree that states and districts should exercise caution when requiring or allowing SPSs in teacher evaluation systems as more research needs to be conducted on SPSs, especially in relation to other measures of teacher effectiveness and in light of any decisions made for accountability-related or summative purposes (Kuhfeld, 2017).

Measurement concerns. One of the greatest measurement concerns regarding SPSs is the likelihood of bias (which subsequently affects reliability and validity). SPS scores can be affected (and thus also potentially deemed unreliable and/or invalid) by a variety of student-, teacher-, and classroom-level factors (Desimone, Smith, & Frisvold, 2010; Driscoll, Peterson, Crow, & Larson, 1985; see Spooren, Brockx, & Mortelmans, 2013). Prior research, albeit mostly in higher education, has shown that both student and teacher demographics and related characteristics (e.g., teacher attractiveness, personality traits) have been significantly associated with teachers' SPS scores. For example, research on SPSs in higher education found that female students tended to give higher ratings to teachers than male students (Basow & Montgomery, 2005; McPherson, Todd Jewell, & Kim, 2009; Smith, Yoo, Farr, Salmon, & Miller, 2007), especially when their teachers were also female (Basow, Phelan, & Capotosto, 2006); students who expected a better grade in a given class tended to give higher ratings than students who expected lower grades (McPherson et al., 2009; Remedios & Lieberman, 2008; Spooren, 2010), and teachers who students deemed as attractive received higher ratings than teachers deemed less attractive (Gurung & Vespia, 2007; Hamermesch & Parker, 2005).

Other potentially biasing effects include the non-random sorting of students into classrooms (which is a noted concern for VAMs especially; Rothstein, 2009, 2010), which can often be related to the student-level demographics just mentioned, as well as a variety of rater biases. Rater biases include the halo effect (see, for example, Pike, 1999), the fatal-flaw effect, selection bias (i.e., not enough students responding to a survey to warrant valid inferences being drawn for a given teacher), social desirability bias (see, for example, Maulana & Helms-Lorenz, 2016), and the central-tendency error (i.e., students being more likely to score teachers using values towards the middle of the scale rather than at the extreme; Popham, 2013), among others.

From a validity standpoint, there are questions as to whether students can objectively rate their teachers (Fauth et al., 2014; Kuhfeld, 2017; Liaw & Goh, 2003). This issue is especially pronounced for younger students, such as those in the elementary grades (De Jong & Westerhof, 2001; Kunter & Baumert, 2006). There is also a concern about students not being able to evaluate all facets of what effective teaching entails. In theory, this concern is alleviated by the use of multiple measures (i.e., classroom observations, in this case) within a teacher evaluation system, but researchers (e.g., Peterson et al., 2000) have noted that survey developers and administrators alike must keep in mind what students can and cannot evaluate. For example, students can assess whether a teacher is engaging and presents new material in interesting ways, but they cannot determine how well a teacher knows content standards (Goe et al., 2008; Peterson et al., 1997). Additionally, since individual student responses are aggregated to the teacher level, an implicit assumption in using SPSs is that the survey data at the teacher level represents the same constructs as the individual student-level responses (Schweig, 2014). If some students interpret SPS items differently from other students, the validity of a teacher's SPS score can be threatened. There have also been discussions about averaging SPS scores across students, as the true variability can easily be perceived as minimal or even non-existent (Kitto, Williams, & Alderman, 2019)

Other concerns also focus on varying levels of validity and reliability across different grade levels and subjects. While some researchers have labeled SPSs as reliable and valid overall, other researchers (e.g., Downer, Stuhlman, Schweig, Martinez, & Ruzek, 2015; Li, 2019; Polikoff, 2015; Sandilos et al., 2019) have cited concerns regarding both issues. For example, Sandilos et al. (2019) found that convergent-related evidence of validity between Tripod survey scores and classroom observation scores was greater in middle school grades than in elementary school grades, and in some cases, there was a negative association between Tripod scores and VAM scores. Lastly, concerns have risen regarding how to interpret SPS scores given the nested nature of students in classrooms/with teachers in schools (Downer et al., 2015; see also Marsh et al., 2012), as methods appropriate for nested or multi-level data have infrequently been used when analyzing SPS data (see, for example, den Brok, Brekelmans, & Wubbel, 2006).

Pragmatic concerns. One of the noted pragmatic concerns regarding SPSs is the difficulty in getting buy-in from teachers, especially when SPSs are used at least in part for summative purposes (Balch, 2016; Schultz et al., 2014). Many teachers fear that students' evaluations would be biased based on how a student personally feels about a given teacher, either overall or on the day of the survey (Kauchak, Peterson, & Driscoll, 1985; Schulz et al., 2014), and SPSs would be, essentially, a popularity contest rather than a valid means by which to evaluate teachers (Fauth et al., 2014). Teachers have also voiced worries about students potentially not taking the surveys seriously, as they might not recognize the potential implications survey results could have for teachers (Nott, 2014; Stecher et al., 2018). Another concern, which can be compounded by any of the previously mentioned issues of bias, is that students might not understand what the survey items are asking (Downer et al., 2015). Regardless of the reason of students' possible poor understanding (e.g., reading comprehension ability, English language skills, specific wording of items), teachers have indicated that they worry that the lack of students' understandings will negatively affect their SPS scores (Stecher et al., 2018).

As mentioned, the breadth and depth of research on SPSs at the elementary and secondary level is lacking, especially in relation to VAM and classroom observation research. Even when significant statistical relationships have been demonstrated between SPSs and other measures of teacher effectiveness, the actual causes behind those relationships remain mostly unknown (Wallace et al., 2016). In addition to more research being needed about SPSs, both overall and in relation to other measures, researchers must also examine whether different subgroups of students (e.g., based on gender, race/ethnicity, SES) in different geographical or situational contexts (e.g., rural versus urban schools, in SE or ELL versus mainstream classrooms) interpret SPS items and SPS constructs consistently. Other more technical details related to SPSs should also not be ignored, such as the method of survey administration (e.g., paper versus online), as these aspects can also affect an SPS's validity and reliability (Kuhfeld, 2017). In sum, SPSs should only be used if and when each specific SPS has been psychometrically validated its intended purpose (AERA, APA, & NCME, 2014; Goe et al., 2008; see also Kane, 2001)—which includes specific grades or developmental levels, student subgroups, geographical/situational contexts, and the like. If SPSs—or any measure of teacher effectiveness, for that matter—are used for purposes other than for which they are intended and/or have been validated, serious implications can result.

Measures Summary

As discussed, each of the three main measures of teacher evaluation used in New Mexico between the 2013-2014 and 2015-2016 school years have their own unique history, development, and use within contemporary teacher evaluation systems. Although much of the teacher evaluation literature over the past decade or two has focused on VAMs, all individual measures of a teacher evaluation system should be critically examined. No one measure is immune to measurement or pragmatic concerns, and creating composite measures and/or using multiple measures within a teacher evaluation system leads to even potentially greater concerns that also must be investigated, especially before such measures or systems are used to inform highly consequently personnel decisions.

One commonality across VAMs, classroom observations, and SPSs is that all three measures should adhere to the *Standards* (AERA et al., 2014) before they are utilized within a teacher evaluation system. Among other criteria, the *Standards* calls for any measure or measurement tool in education to meet several methodological standards to ensure that each measure is used appropriately. Three of these standards—reliability, validity, and (a lack of) bias—form the basis of the conceptual framework in which I used to situate this study. In the following section, I explain my conceptual framework and discuss each of these three standards in turn.

Conceptual Framework

The conceptual framework undergirding this study is drawn from the *Standards for Educational and Psychological Testing* (AERA et al., 2014). Specifically, I framed this study using key measurement concepts that the AERA, APA, and NCME outline as integral to educational measurement: reliability, validity, and (a lack of) bias.

While there are additional concepts that are of high importance to educational measurement and testing (e.g., fairness, transparency, appropriate use), reliability, validity, and bias are the three key concepts that should be assessed not just from a pragmatic or applied standpoint, but from a measurement one as well. In other words, if measures of teacher effectiveness are unreliable, invalid, and/or biased, the inferences drawn from them are inherently flawed, regardless of whether those measures are fair, transparent, and properly used. The below subsections outline each of these three essential measurement concepts, both in general but specifically to VAMs.

Reliability

In the context of teacher effectiveness measures, reliability is the degree to which test- or measurement-based scores "are consistent over repeated applications of a measurement procedure [e.g., a VAM] and hence are inferred to be dependable and consistent" (AERA et al., 2014, pp. 222-223) for the individuals (e.g., teachers) to whom the test- or measurement-based scores pertain. Specific to teacher effectiveness measures, reliability (i.e., intertemporal stability; see, for example, McCaffrey et al., 2009) should be observed when estimates of each measure of teacher effectiveness are more or less consistent over time, from one year to the next, regardless of the compositions of students within a teacher’s classes or within a teacher’s school. This consistency is typically measured using statistics like standard errors, reliability and generalizability coefficients, or other markers of classification consistency.

Within the teacher evaluation literature, reporting on the reliability of measures is necessary to make transparent the potential lack of consistency over time for a given measure, either overall or in specific contexts. This transparency allows for researchers and the potential users of a given measure to understand the contexts in which inferences from the given measure are made, which is especially critical when a measure is used (either in isolation or in combination with other measures) to inform high-stakes decisions. Further, without adequate reliability, it is nearly impossible to defend the validity and proper use of a measure (Brennan, 2013; Kane, 2006, 2013; Messick 1975, 1980).

When a teacher effectiveness measure is determined to be unreliable (i.e., inconsistent over time), the result can lead to a teacher being incorrectly classified. While both false positives (i.e., an ineffective teacher being classified as an effective teacher; a Type I error) and false negatives (i.e., an effective teacher being classified as an ineffective teacher; a Type II error) warrant concern, false positives are especially concerning since “failing to identify teachers who are truly ineffective poses risks to students” (Raudenbush & Jean, 2012, p. 12).

Out of the common teacher effectiveness measures, VAMs appear to be more inconsistent than other measures. Researchers have found that the likelihood of a teacher being misclassified per their VAM estimate can range from 25% to as high as 59% (Martinez et al., 2016; Schochet & Chiang, 2013; Yeh, 2013). While reliability can be increased with three years of data, there still exists at least a 25% chance that teachers may be misclassified. Additionally, after including three years of data, the strength that additional years of data add to the reliability of VAM estimates plateaus (Brophy, 1973; Cody, McFarland, Moore, & Preston, 2010; Glazerman & Potamites, 2011; Goldschmidt, Choi, & Beaudoin, 2012; Harris, 2011; Ishii & Rivkin, 2009). The likelihood that one out of every four teachers might be incorrectly classified by VAMs is noteworthy enough to warrant caution, especially before high-stakes consequences are attached to VAM output (see, for example, Briggs & Domingue, 2011; Chester, 2003; Glazerman et a., 2011; Guarino, Reckase, & Wooldridge, 2012; Harris, 2011; Rothstein, 2010; Shaw & Bovaird, 2011; Yeh, 2013).

Validity

In the context of teacher effectiveness, validity is “the degree to which evidence and theory support the interpretations” of the various scores of each measure used to evaluate teachers, per each measure’s proposed use (AERA et al., 2014, p. 11). AERA et al. specifically notes that a measure alone cannot be defined as valid without qualifying statements about the interpretation of scores and its proposed use(s). When establishing evidence of validity, one must be able to provide evidence that accurate inferences can be drawn from the data for whatever inferential purposes the data are being used (see Cronbach & Meehl, 1955; Kane, 2006, 2013).

While there are multiple types of validity (e.g., content-related, criterion-related, construct-related, consequential-related), the most often examined validity regarding teacher effectiveness is convergent-related evidence of validity. In this context, convergent-related evidence of validity is the degree of the relationship between two different measures of teacher effectiveness that have been taken at the same time (Messick, 1989). This type of validity is important to establish as it is used to assess the extent that different measures of similar constructs converge. In this case, the overall construct that each individual measure is trying to asses is “teacher effectiveness.” Gathering evidence of convergent-related evidence of validity is necessary to determine whether teachers who are deemed effective (or ineffective) by one measure are also deemed effective (or ineffective) by other measures that are collected at the same time in the same contexts.

In terms of teaching effectiveness measures, some argue that any indicators (i.e., measures) that are mapped onto the general construct of “teaching effectiveness” should have a strong relationship. Others, conversely, argue that a weak relationship (i.e., a low correlation) between any two measures tells us nothing about whether either one, neither, or both are useful. Specific to VAMs, research evidence suggests that VAM estimates of teacher effectiveness have a relatively weak relationship (i.e., do not strongly correlate) with other common measures, such as classroom observation scores. While some argue that the non-VAM measures are more to blame for these poor relationships, others argue that all of the measures are to blame, including VAMs, because they are all highly imperfect and flawed (Gabriel & Lester, 2013a, p. 4; see also Harris, 2011).

Regardless of which measure is more “at fault” for the poor relationship, research has typically demonstrated that the correlations between VAM scores and classroom observation scores or student surveys, respectively, are low to moderate, at best. (Grossman, Cohen, Ronfeldt, & Brown, 2014; Harris, 2011; Hill et al., 2011; see also Koedel et al., 2015). These relatively weak correlations are also akin to those observed via the aforementioned MET study in which researchers searched for and assessed the same evidences of convergent-related validity (Kane & Staiger, 2012; see also Polikoff & Porter, 2014; Rothstein & Mathis, 2013).

While the actual values that quantify the relationship between VAMs and other measures have been, overall, relatively weak, different researchers and scholar have interpreted these values differently. That is, some (e.g., typically the proponents of VAMs) interpret these values as high enough that they can conclude that convergent-related evidence of validity has been demonstrated or as high enough to support one or both measures’ uses within teacher evaluation systems. However, others (e.g., typical those who are against VAMs, or against VAMs for summative purposes) interpret these values as being much too low to reach those same conclusions, and instead interpret such low values as a warning signal that one or both measures should not be used, especially for high-stakes purposes.

Bias

In the context of teacher effectiveness, bias occurs when teachers' scores for a given measure vary based on characteristics that are not relevant to the measure itself. Put differently, bias exists when a "student, teacher, or course characteristic affects [a teacher's effectiveness score], either positively or negatively, but is unrelated to any criteria of good teaching" (Centra, 2003, p. 498). A more technical definition of bias is the "construct underrepresentation of construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers and consequently the…validity of interpretations and uses of their test scores" (AERA et al., 2014, p. 216). In this context, bias is observed if a measure of teaching effectiveness is significantly correlated with, for example, student demographic variables. Put differently, for example, a teacher with larger proportions of Hispanic students should not consistently receive lower scores on a given measure than a teacher with smaller proportions of Hispanic students. When a measure is highly correlated with potentially biasing factors, it is no longer possible to make valid inferences about that measure's score as evidence of bias results in measures' interpretations being distorted (Messick, 1989; see also Haladyna & Downing, 2004).

Regarding teacher effectiveness measures, as previously discussed, all are susceptible to potentially biasing factors. Most of the concern regarding bias has been with VAMs, given the propensity for VAMs to factor most heavily in a teacher's overall evaluation. Over the past decade VAM-based evidence of bias has been investigated at least 33 times in articles published in top peer-reviewed journals (Lavery, Amrein-

Beardsley, Geiger, & Pivovarova, in press). Evidenced across these articles is that bias is still of great debate, as is whether statistically controlling for bias by using complex statistical approaches (e.g., VAMs) to account for non-random student assignment makes such biasing effects negligible or “ignorable” (Rosenbaum & Rubin, 1983; see also Chetty et al., 2014a, 2014b; Koedel et al., 2015; Rothstein, 2017).

In sum, as per the *Standards* (AERA et al., 2014), ongoing evaluation of these three measurement issues as pertaining to all teacher effectiveness measures and their uses is essential. However, while essential, the thoroughness of the research of these measures is also critical. This was the case with VAM research, to the point where the ASA (2014), the AERA Council (2015), and the National Academy of Education (Baker et al., 2010) have underscored similar calls for research within their associations’ positions statements about VAMs and VAM use (see also Harris & Herrington, 2015). While no measure of teacher effectiveness is perfect or without error, each and every measure used, regardless whether for formative or summative purposes, should aim to be as reliable, valid, and unbiased as possible.

CHAPTER 3

METHODOLOGY

In this chapter, I describe the methodology I used in this study to answer my two overarching research questions about the relationships among measures of teacher effectiveness and the effects of different student compositions within a teacher's school. The research questions were: 1) What are the relationships between student background characteristics, aggregated to the school level, and the four main teacher evaluation measures that comprised a teacher's overall evaluation score in New Mexico during the 2013-2014, 2014-2015, and 2015-2016 school years? and 2) How do these relationships compare across the four main teacher evaluation measures? To answer these two questions, I answered the following sub-questions, grouped by teacher evaluation measure:

1. VAS scores:

   1a.) What is the relationship between the percent of special education (SE) students within a teacher's school and the percent of VAS points a teacher earns?

   1b.) What is the relationship between the percent of English language learner (ELL) students within a teacher's school and the percent of VAS points a teacher earns?

   1c.) What is the relationship between the percent of students eligible for free and reduced lunch (FRL) within a teacher's school and the percent of VAS points a teacher earns?

1d.) What is the relationship between the percent of underrepresented minority (URM) students within a teacher's school and the percent of VAS points a teacher earns?

2. Classroom observation scores:

2a.) What is the relationship between the percent of SE students within a teacher's school and the percent of observation points a teacher earns?

2b.) What is the relationship between the percent of ELL students within a teacher's school and the percent of observation points a teacher earns?

2c.) What is the relationship between the percent of FRL students within a teacher's school and the percent of observation points a teacher earns?

2d.) What is the relationship between the percent of URM students within a teacher's school and the percent of observation points a teacher earns?

3. Planning, Preparation, and Professionalism (PPP) scores:

3a.) What is the relationship between the percent of SE students within a teacher's school and the percent of PPP points a teacher earns?

3b.) What is the relationship between the percent of ELL students within a teacher's school and the percent of PPP points a teacher earns?

3c.) What is the relationship between the percent of FRL students within a teacher's school and the percent of PPP points a teacher earns?

3d.) What is the relationship between the percent of URM students within a teacher's school and the percent of PPP points a teacher earns?

4. Student Perception Survey (SPS) scores:

4a.) What is the relationship between the percent of SE students within a teacher's school and the percent of SPS points a teacher earns?

4b.) What is the relationship between the percent of ELL students within a teacher's school and the percent of SPS points a teacher earns?

4c.) What is the relationship between the percent of FRL students within a teacher's school and the percent of SPS points a teacher earns?

4d.) What is the relationship between the percent of URM students within a teacher's school and the percent of SPS points a teacher earns?

As previously discussed in Chapter 2, while the prior work taken on by Amrein-Beardsley and Geiger (revise and resubmit) illuminated some of the potential issues with New Mexico's VAS data, they failed to control for a variety of factors that might have influenced the significance of their results. Thus, via this study I aim to expand on their work and fill that gap by analyzing the VAS data with taking those factors into account (see details forthcoming). In the following sections, I first explain the analytic method I used to answer my research questions. I then discuss my data source and participants, including my sampling procedure that resulted in the final data sets that I utilized for these analyses. Lastly, I conclude with a discussion of study limitations.

Data and Participants

Data Source

I acquired data for this study from the aforementioned *State ex rel. Stewart v. New Mexico Public Education Department* (2015) lawsuit. The NMPED was required to provide the plaintiffs' lawyers with New Mexico's 2013-2014, 2014-2015, and 2015-2016 teacher evaluation data for expert witness analyses (as described above; see also Amrein-Beardsley & Geiger, revise and resubmit). It is important to note that, for this study, Arizona State University (ASU) Institutional Review Board (IRB) approval was not needed as I was using secondary data. I did, however, receive permission from the plaintiffs' lawyers in the above-mentioned lawsuit to use the data for this study (see Appendix A).

The NMPED provided six data files that contained data on all public and charter school teachers in the state. Per academic year, there was one file with teacher evaluation data and one file with aggregated student-level classroom composition data. The teacher evaluation data files contained teacher-level demographics (e.g., age, gender, years of experience), position information (e.g., staff status, school level taught, teacher title), and teacher evaluation scores (e.g., percent of VAS points earned, possible observation points and observation points earned, summative evaluation score). The aggregated student-level classroom composition files contained percentages of student subpopulations per classroom (e.g., percent of students per class by race/ethnicity, SE status, FRL status). All files contained teacher license numbers, and I used those numbers to aggregate all classroom-level data per teacher, and then link that data to the teacher evaluation data.

Data Population and Sampling

There were 29,967 unique teachers across the entire dataset (i.e., from the 2013-2014 through 2015-2016 school years). To narrow down the dataset to my final sample, I used several inclusion criteria, per year. First, and per year, each teacher had to have been evaluated with all four measures under analysis in this study (i.e., VAMs, classroom observations, the PPP measure, and student surveys). I coded each teacher as being evaluated with all four measures if s/he did not have any missing data for each of the four measures. Barring any odd error in the data file (e.g., a teacher being incorrectly linked to another teacher's evaluation scores), this inclusion criterion would essentially guarantee that each teacher was evaluated by each of the four measures. Additionally, I wanted to avoid the problem of missing data as much as possible. While there are multiple ways to statistically account for missing data, (e.g., multiple imputation, controlling for missing values by using dummy variables, pairwise or casewise deletion) many of these methods can result in potentially serious errors that can introduce substantial bias in an analysis (Enders, 2010), which can subsequently result in inaccurate inferences. Further, generally speaking, it is possible that teachers who are evaluated by different sets of measures are inherently different from each other, in terms of the grades, subjects, or types of students they teach. I wanted my sample to be as homogenous as possible in this regard to reduce the potential for bias.

Second, each teacher had to be employed at a public school (i.e., not a charter school). I excluded all charter school teachers as, typically, charter schools are often very different than traditional public schools in many aspects (e.g., governance, student enrollment, personnel, funding, accountability, curriculum) (Lubienski, 2002; Podgursky & Ballou, 2001; Shober, Manna, & Witte, 2006). This is the case in New Mexico, where charter schools operate differently (e.g., governance, funding) and have different school characteristics (e.g., student body composition, student-teacher ratio) than traditional public schools (Charter Schools Act, 2007; see also NMPED, 2016b). Although charter schools utilized the NMTEACH system for teacher evaluation (NMPED, 2016a), due to the differences noted above, I excluded all charter school teachers from the sample.

All teachers in the final sample also needed to have course-specific data (i.e., data on the students that the teachers taught). I wanted to aggregate individual teachers' students to the teacher level (i.e., across all courses a given teacher taught, as many teachers taught more than one course) as teachers' overall VAS scores were reflective of all of their students. Additionally, research on classroom observations indicates that all of a teacher's classes should be taken into consideration for evaluation purposes (e.g., for improved validity and reliability; Lei, Li, & Leroux, 2018). Thus, it was imperative that these course data be present. Related, each teacher needed to have taught at least 15 students across all courses, per year, to ensure improved reliability for VAM estimates. McCaffery et al. (2009) explained how, in their study of year-to-year variability of value-added estimates, teachers who taught under 15 total students had imprecise and unstable estimates, and inflated standard errors. In their study, using a 15-student threshold led to more reliable and precise estimates, and their sample size remained large enough for their analyses. Lastly, all teachers needed to be classified as "certified personnel" as this ensured that all teachers in the sample received comparable education and training (i.e., compared to non-certified teachers). Restricting the sample to those classified as such ensured that potential differences in teacher evaluation scores could not be attributed to a potential lack of certification.

These inclusion criteria resulted in a final sample across all years of 10,686 unique teachers. Following the above inclusion criteria, there were 2,733 unique teachers in Year 1, 2,738 unique teachers in Year 2, and 8,963 unique teachers in Year 3. The large increase in the Year 3 sample is due to the overall number of teachers who had SPS data. Between the 2013-2014 and 2015-2016 school years, each district was allowed to individually decide whether to use SPSs to evaluate teachers (see Doan et al., 2019). Although not explicitly stated, I surmise that between 2014-2015 (Year 2) and 2015-2016 (Year 3), the number of districts using SPSs dramatically increased. In Year 1 and Year 2, 27% (*n* = 5,692/20,677) and 26% (*n* = 5,616/21,427) of all teachers present per year, respectively, had SPS data. In Year 3, that proportion jumped to 59% (*n* = 12,531/21,140) of teachers having SPS data, which explains the increase in the number of teachers in the Year 3 sample.

Analytic Plan

Multiple Linear Regression

To answer my research questions, I utilized multiple linear regression. In general terms, linear regression is most useful when wanting to determine what effect, if any, one or more variables (i.e., predictor variables, or independent variables [IVs]) have on another variable (i.e., the criterion variable, or dependent variable [DV]). Typically, both the predictor variables and the criterion variable are continuous, but the predictor variables do not have to be. Unlike the Pearson product-moment correlation coefficient (which is frequently used when wanting to determine the relationship between two variables), which simply provides a value that quantifies the relationship between two variables, linear regression can be utilized to evaluate how well one variable can predict a second variable (Lomax & Hahs-Vaughan, 2012). Additionally, if there is a significant relationship between two variables, regression allows for examining both the strength and direction of that relationship (Keith, 2015). Linear regression with only one predictor variable is known as simple linear regression, where linear regression with two or more predictor variables is known as multiple linear regression.

While simple linear regression might be useful to determine the predictive relationship between two variables, it does not allow for additional predictor variables (i.e., covariates) to be included, which is a large drawback since "most phenomena of interest have multiple causes" (Berry & Sanders, 2000, p. 32). Including additional predictor variables (i.e., covariates) in a regression model is necessary when those additional variables are believed (e.g., from theory) or known (e.g., from prior research) to have a relationship with the criterion variable (Stock & Watson, 2007). Multiple regression provides a way to assess the effect of a single predictor variable on the criterion variables while holding all other predictor variables constant (i.e., controlling for all other predictor variables) (Berry & Sanders, 2000).

In multiple linear regression, the coefficient of multiple determination (i.e., R^2) is used to quantify how much variance in the criterion variable is explained by the predictor variables in a given model. R^2 values range from 0 to 1, and higher values of R^2 indicate that the predictor variables in a model have more "explanatory power" (Berry & Sanders, 2000, p. 45). For example, an R^2 of .5000 would indicate that 50% of the variance in the criterion variable could be explained by the cumulative effect of the predictor variables in a given model. Unlike other types of model fit statistics more commonly used in other statistical procedures (e.g., structural equation modeling), there is no “gold standard” that defines an R^2 as large enough to categorize the predictor variables in a model as explaining a “meaningful” percent of variance in the criterion variable (Lomax & Hahs-Vaughn, 2012, p. 376). This is because an R^2 value is affected by multiple aspects of a model (e.g., number of predictor variables, quality of predictor variables, variation in the criterion variable). For example, R^2 typically increases (albeit possibly nominally) solely from adding an additional predictor variable to a model, regardless of the strength of the relationship of that predictor variable to the criterion variable (Pelham, 2013).

Lastly, it is also important to note that in choosing multiple linear regression as my method, I am making several assumptions, all of which are standard assumptions of multiple linear regression (see Berry & Sanders, 2000; Osborne & Waters, 2002). First, I assumed a linear relationship between each predictor variable and criterion variable. I also assumed that the “effects of all [predictor] variables on the [criterion] variable are additive” (Berry & Sanders, 2000, p. 38). In other words, I assumed that there are no interaction effects between predictors variables on the criterion variable. Lastly, and possibly most importantly, I assumed that all variables were measured without error. It is important to note that it is not possible to “confidently claim that *all* assumptions…have been satisfied completely” (Berry & Sanders, 2000, p. 24), as many assumptions are impossible to empirically test. Rather, “whether an assumption has been met is really a question of degree” (p. 24).

Creation of Selected Predictor Variable, Covariates, and Criterion Variables

The data files provided by the NMPED did not contain all of the necessary variables needed for the models I wanted to run, so in several cases I had to clean, transform, or aggregate the provided data so it would be suitable for analyses. The following subsections detail these processes.

Predictor variable and covariates. The predictor variable that I had to derive was the percent of URM students in a teacher's school, and the two covariates I had to derive were the percent of URM students in a teacher's school and teacher URM status. The NMPED data files provided the counts and percentages of students within a teacher's classes and within a teacher's school, respectively, broken out by individual race/ethnicities, along with each teacher's race/ethnicity. It should be noted that race/ethnicity was reported as a categorical variable without a multi-select option. That is, each teacher and the student counts were identified by one race/ethnicity only. Related, it was not clear how these identities were created (i.e., did teachers/students self-identify, were they ascribed a racial/ethnic category by some other means, or another method entirely). To create my three URM variables, I defined URM as any race or ethnicity other than Caucasian or Asian (i.e., African American, Hispanic, Native American), which is a common way to categorize URMs in the state of New Mexico and states with similar demographics (see, for example, New Mexico Alliance for Minority Participation, 2016). If a teacher had a missing value for race/ethnicity, s/he was coded as non-URM as I could not verify his/her URM status.

It should be noted that a teacher's race/ethnicity and gender were missing in the Year 2 data file from the NMPED. If a teacher was present in Year 1 and/or Year 3, I was able to fill in the missing race/ethnicity and gender values for Year 2. If a teacher was not present in either Year 1 and/or Year 3, I coded that teacher's race/ethnicity and/or gender as missing. As noted above, the missing data for race/ethnicity affected the values of the teacher URM variable. Lastly, as relevant to using Year 1 or Year 3 race/ethnicity or gender values for Year 2, I treated both of these constructs as static and unchanging from year to year, even though this does not have to be the case for either datapoint (Bem, 1993, 1995; Cornell, 1996; Nagel, 1994; Sweetnam, 1996).

Criterion variables. The three criterion variables I had to derive were the percent of observation points, the percent of PPP points, and the percent of SPS points. These derivations were dictated by teachers' VAS scores, as VAS score data provided by the NMPED were represented as the percent of VAS points earned out of the total possible VAS points. For comparison purposes, I wanted each of the four criterion variables to be of the same data type, so I derived the three aforementioned variables by dividing each teacher's earned points by possible points, per measure.

Regression Models

To determine what, if any, bias in teachers' four measures of evaluation existed from the student compositions within their schools, I conducted 48 multiple linear regressions using Stata 14.2 SE (StataCorp, n.d.). For all models, my null hypothesis was that the main predictor variable of interest (along with the specified covariates in the model) had no effect on the criterion variable (i.e., the percent of points earned for a given measure of teacher effectiveness) (α = .05). I ran separate models per year of data in my dataset (discussed in more detail, below) for all four teacher evaluation measures (i.e., VAS scores, classroom observation scores, PPP scores, SPS scores) discussed above. Thus, I had one model per year per evaluation measure per predictor variable. In other words: [four criterion variables] x [four main predictor variables] x [three separate years of data] = 48 regressions.

In each regression, I controlled for a variety of teacher-, class- and school-level factors. As mentioned above, it is standard practice to include variables in a model that might be correlated with the criterion variable (other than the main predictor variable of interest) to ensure the statistical association between the main predictor variable of interest and the criterion variable cannot be explained due to other omitted factors (Pelham, 2013; Stock & Watson, 2007). For example, in the models where the criterion variable was the percent of teachers' classroom observation scores, I controlled for teachers' years of experience (i.e., a predictor variable that is likely correlated with teachers' years of experience) as prior research (e.g., Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2008; Ladd & Sorensen, 2017; see also Darling-Hammond, 2000; Podolsky, Kini, & Darling-Hammond, 2019) and logic (e.g., Becker, 1964) would indicate that teachers with more years of experience might have higher classroom observation scores. However, since this study was not a true experiment (i.e., as students were not randomly assigned to teachers/schools; teachers were not randomly assigned to schools/districts), there were likely unaccounted influences affecting the criterion variables in the models (see Berliner, 2014). Thus, any predictor variable that has a significant association with any of the criterion variables should be interpreted with caution. That is, the significant effect of a predictor variable on a criterion variable should not be interpreted as casual (Angrist & Pischke, 2009). See Appendix B for specifics about each of the 48 regression models, including the criterion variable of interest, the main predictor variable of interest, and covariates.

Study Limitations

Like all research, this study was not without limitations. One of the biggest limitations was not having full model specifics (i.e., the mode's source code) for the VAM that was used to calculate New Mexico teacher's VAS scores. Without knowing this information, I was not able to replicate the exact model used to derive teachers' VAS scores. Having access to this information would have allowed me to replicate the actual model and causally test to what extent student background characteristics did or did not bias teachers' VAS scores (and, subsequently, teachers' overall evaluation scores). Without this information, no true causal statements can be made about the associations between student background characteristics and teachers' VAS scores. It is also worth noting that the lack of transparency into VAM model specifics is in fact a limitation of nearly all VAM-related research, due to the proprietary nature of VAMs, save for those studies conducted by researchers who are affiliated with and presumably have access to both the model itself (i.e., what variables are included in a given VAM) and raw student-level data (e.g., test scores, demographics) (e.g., Sanders & Horn, 1994, 1998; Sanders et al., 1997; Wright, White, Sanders, & Rivers, 2010).

Second, the data files provided by the NMPED did not contain any raw data at the student level, such as individual students' academic data (e.g., test scores) or demographic factors[4] (e.g., income), for example. It is unknown if such data were missing because the NMPED simply did not include them in the files sent to the plaintiffs' lawyers, or if the NMPED did not collect such data, in general. The lack of data at the student level—especially the demographic factors—is a severe drawback. This drawback is actually the case in most research that focuses on student achievement data, as prior research has indicated that non-school factors (e.g., poverty, SES) are stronger drivers of student achievement than in-school factors (e.g., Berliner, 2009; Coleman et al., 1966; Jencks et al., 1979; Hanushek et al., 2003). Since student achievement data is a main component of teachers' VAS scores and VAS scores constitute a large portion of teachers' evaluation scores, and since prior research has called into question the extent to which VAMs adequately and accurately control for such out-of-school factors (e.g., Amrein-Beardsley & Holloway, 2019; Ishii & Rivkin, 2009; Kupermintz, 2003; Scherrer, 2011; Tekwe et al., 2004), this study would have potentially been greatly strengthened if the NMPED had included such data in their files.

Third, it is possible that missing values in the data files from the NMPED might have affected the significance of certain model results. For example, the school level a teacher taught was not present in the Year 3 file, so it was not possible to include that data point as a control variable in the 16 Year 3 models. It is possible that if that variable

[4] The demographic factors that were included (e.g., percent of SE students, percent of FRL students) were not at the student level, but rather were aggregated to either the teacher level across all of a teacher's classes or the school level.

had been present in the data files and thus, had been included in the 16 Year 3 models, the coefficients or the significance of the main and/or additional predictor variables in those models would have been altered. It is also possible that the teachers for which certain data were missing, which resulted in those teachers being excluded from the study, were inherently different from those teachers for whom data was not missing. It is not possible to test this potential difference between groups of teachers, as I was not able to conclude whether such missing data was missing completely at random or missing not at random (Enders, 2010). However, it is possible that if currently excluded teachers had been included in the study, coefficients, model significance, and overall conclusions might have changed.

A last limitation, and one that has been pointed out by peer reviewers when reviewing the prior analyses completed by Amrein-Beardsley and Geiger (revise and resubmit), is that this study was limited to teacher evaluation measures in one state. The general population of New Mexico—in terms of both children (i.e., students) and adults (i.e., teachers) is quite different from the populations in other states (e.g., based on race/ethnicity, socioeconomic status, state funding for education), so any potential findings cannot necessarily be generalized to other states. Related, the entire state of New Mexico only uses one VAM (see Martinez et al., 2016; see also Reiss, 2017), one observational framework (Danielson's *Framework for Teaching*; The Danielson Group, 2013), and one student survey, which maps onto three of Danielson's four domains (a modified Tripod survey; see Ferguson, 2008; NMPED, n.d.g), for its teacher evaluation scores. Also from a generalizability standpoint, it is possible that analyzing teachers of identical demographics and identical teaching effectiveness in other states, within a different evaluation system, or using different evaluation measures—either different forms of the same measures or different measures entirely—might produce entirely different results. As such, findings from this study should be understood in the context of these measures only, and not be explicitly generalized to other VAMs, observational frameworks, or student surveys. Findings should also not be generalized outside of the state of New Mexico, or outside of the 2013-2014, 2014-2015, and 2015-2016 school years within the state of New Mexico.

CHAPTER 4

RESULTS

In this chapter, I present the findings from my 48 regression models, which I used to answer the following research questions detailed prior and restated here: 1) What are the relationships between student background characteristics, aggregated to the school level, and the four main teacher evaluation measures that comprised a teacher's overall evaluation score in New Mexico during the 2013-2014, 2014-2015, and 2015-2016 school years? and 2) How do these relationships compare across the four main teacher evaluation measures?

I first describe the teachers in my sample, overall and then per year. I then detail the findings for each regression model, organized first by teacher evaluation measure (i.e., the criterion variable in each model) and then by year. I then summarize the results across each teacher evaluation measure.

Sample Demographics

Full Sample

Across all teachers present in the full sample (*n* = 10,686), the majority were female (61%; *n* = 6,562/10,686) and non-URM (62%; *n* = 6,640/10, 686), with an average of 11.3 years of experience (*SD* = 9.46). Most teachers were also classified as regular classroom teachers (87%; *n* = 9,321/10, 686), with few being classified as Special Education teachers (7%; *n* = 718/10, 686) and even fewer being classified as bilingual teachers[5] (4%; *n* = 418/10, 686) (see Table 3 for full demographic data, overall and per year).

[5] Being a regular classroom teacher, Special Education teacher, and/or bilingual teacher were not mutually exclusive classifications per the NMPED data.

Table 3

Teacher Demographics, Overall and Per Year

	Across All Years		Year 1		Year 2		Year 3	
			2013-2014		2014-2015		2015-2016	
	n	Pct.	n	Pct.	n	Pct.	n	Pct.
Teachers	10,686	100%	2,733	100%	2,738	100%	8,963	100%
Gender								
Female	6,562	61%	2,146	79%	1,934	71%	5,311	59%
Male	2,081	19%	587	22%	658	24%	1,684	19%
Missing	2,043	19%	0	0%	146	5%	1,968	22%
Race/Ethnicity								
Caucasian	6,364	60%	1,543	57%	1,439	53%	5,352	60%
Hispanic	3,601	34%	1,040	38%	1,078	39%	3,055	34%
Native Am.	316	3%	86	3%	88	3%	249	3%
African Am.	129	1%	22	1%	26	1%	113	1%
Asian	227	2%	42	2%	58	2%	194	2%
Missing	49	0.5%	0	0%	49	2%	0	0%
Underrepresented Minority (URM)								
Yes	4,046	38%	1,148	42%	1,192	44%	3,417	38%
No	6,640	62%	1,585	58%	1,546	56%	5,546	62%
Regular Classroom Teacher								
Yes	9,321	87%	2,352	86%	2,434	89%	8,020	89%
No	1,365	13%	381	14%	304	11%	943	11%
Special Education (SE) Teacher								
Yes	718	7%	201	7%	143	5%	551	6%
No	9,968	93%	2,532	93%	2,595	95%	8,412	94%
Bilingual Teacher								
Yes	418	4%	157	6%	156	6%	341	4%
No	10,268	96%	2,576	94%	2,582	94%	8,622	96%
Total Years of Experience								
0-2 Years	2,250	21%	437	16%	483	18%	1,986	22%
3-8 Years	2,571	24%	657	24%	635	23%	2,186	24%
9-15 Years	2,628	25%	766	28%	721	26%	2,208	25%
16+ Years	3,170	30%	858	31%	899	33%	2,517	28%
Missing	67	1%	15	1%	0	0%	66	1%
Total Years of Experience	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
	11.3	9.46	12.2	9.46	12.2	9.59	10.9	9.04

Note: Percentages might not add to 100 due to rounding.

Of the 10,686 teachers, 73% (*n* = 7,757/10,686) were present in all three sample years (i.e., 2013-2014, 2014-2015, and 2015-2016). Another 494 teachers were present in Years 1 and 2 only (5%; *n* = 494/10,686), and another 1,114 teachers were present in Years 2 and 3 only (10%; *n* = 1,114/10, 686). The remaining teachers were present in Years 1 and 3 only (1%; *n* = 116/10, 686), Year 1 only (3%; n = 306/10, 686), Year 2 only (0.3%; n = 28/10, 686), or Year 3 only (8%; n = 871/10, 686).

Year 1 (2013-2014) Sample

Of teachers present in Year 1, 79% were female (*n* = 2,146/2,733)[6], 58% were non-URM (*n* = 1,585/2,733), and 86% were regular classroom teachers (*n* = 2,352/2,733). Only 7% and 6% were classified as SE teachers (*n* = 201/2,733) or bilingual teachers (*n* = 157/2,733), respectively. On average, teachers had 12.2 years of experience (*SD* = 9.46) (see Table 3, again).

Teachers taught an average of 1.8 classes and 73 students across all of their classes (see Table 4). The majority of students were male (52%), qualified for FRL (71%), and were URM (74%). Sixteen percent of students were labeled as SE, and 5% of students were labeled as gifted. The percent of ELL students per class were not listed in the 2013-2014 dataset from the NMPED, so it was not possible to report on the percent of ELL students in the classes of teachers in the Year 1 sample.

[6] Year 1 was the only year with no missing values for gender. Five percent of Year 2 teachers had missing data for gender and 22% of Year 3 teachers had missing data for gender. The percentages of female (and male) teachers in Year 1 is affected by the lack of missing data as there is one fewer categorical option for this variable in Year 1.

Table 4

Teacher-Level Classroom Demographics, per Year

	Year 1 2013-2014			Year 2 2014-2015			Year 3 2015-2016		
	M	SD	Range	M	SD	Range	M	SD	Range
Total Courses per Teacher	1.8	1.29	1 – 13	2.4	1.52	1 – 15	2.3	1.44	1 – 21
Total Students per Teacher	72.8	67.64	15 – 1,424	73.1	53.08	15 – 622	83.4	63.56	15 – 555
Percent of Male Students	52%	9.3%	13% - 100%	51%	10.1%	6% - 95%	51%	9.7%	0% - 100%
Percent of SE Students	16%	22.4%	0% - 100%	14%	18.3%	0% - 100%	15%	19.8%	0% - 100%
Percent of Gifted Students	5%	10.0%	0% - 100%	5%	8.3%	0% - 100%	6%	10.3%	0% - 100%
Percent of FRL Students	71%	27.7%	0% - 100%	76%	28.1%	0% - 100%	75%	27.6%	0% - 100%
Percent of URM Students	74%	21.9%	0% - 100%	76%	22.5%	7% - 100%	74%	21.6%	0% - 100%

The distribution of teachers' overall teacher evaluation ratings was normal (i.e., a bell curve), with the majority of teachers earning a score of "Effective" (48.4%; *n* = 1,324/2,733) fewer and similar percent of teachers earning scores of "Minimally Effective" (28.5%; *n* = 778/2,733) and "Highly Effective" (17.3%; *n* = 474/2,733), respectively; and even fewer teachers and similar percent of teachers earning scores of "Ineffective" (4.5%, *n* = 122/2,733) and "Exemplary" (1.3%, *n* = 35/2,733) (see Figure 3).

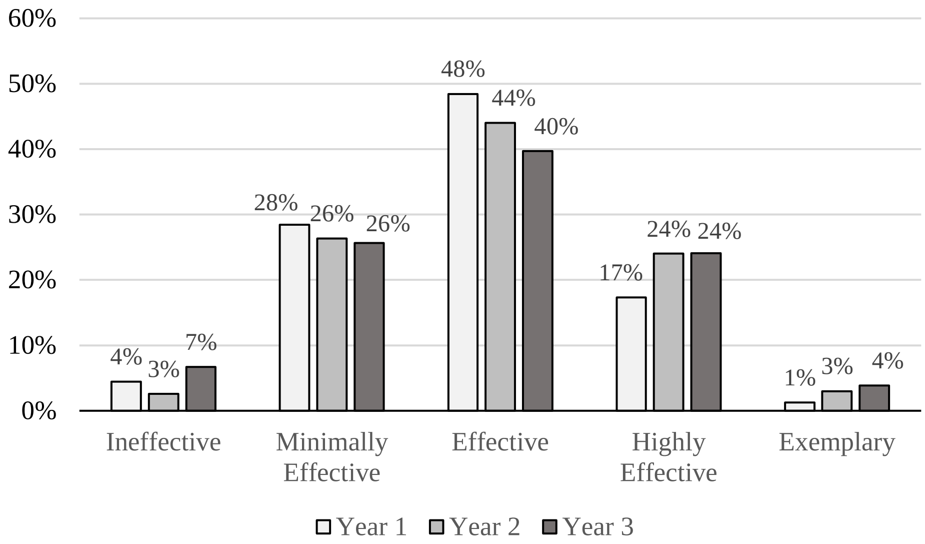


Figure 3. Distribution of teachers' evaluation ratings, per year.

In line with teachers' overall ratings, the average summative score was a 2.8 (*SD* = 0.81). On average, teachers in the Year 1 sample earned just over half of the possible VAS points (*M* = 0.51, *SD* = 0.231), two thirds of the possible classroom observation points (*M* = 0.67, *SD* = 0.094), nearly 70% of the possible PPP points (*M* = 0.69, *SD* = 0.106), and just over three quarters of the possible SPS points (*M* = 0.76, *SD* = 0.133) (see Table 5).

Table 5

Teacher Evaluation Measures, per Year

	Year 1 2013-2014			Year 2 2014-2015			Year 3 2015-2016		
	M	SD	Range	M	SD	Range	M	SD	Range
Summative Score	2.8	0.81	1 – 5	3.0	0.85	1 – 5	2.9	0.93	1 – 5
Percent of VAS Points Earned	51%	23.1%	0% - 100%	53%	24.0%	0% - 100%	53%	24.5%	0% - 100%
Percent of Observation Points Earned	67%	9.4%	28% - 100%	70%	10.0%	35% - 100%	72%	9.9%	27% - 100%
Percent of PPP Points Earned	69%	10.6%	13% - 100%	73%	11.0%	23% - 100%	73%	11.4%	20% - 100%
Percent of SPS Points Earned	76%	13.3%	0% - 100%	82%	10.1%	23% - 100%	81%	10.6%	4% - 100%

Note: Summative scores range from 1 to 5. Each of the four measures of teacher evaluation are presented as the percent of total points earned.

The distribution of the percent of VAS points earned was relatively normal, while the other three measures were more negatively skewed (see Figure 4).

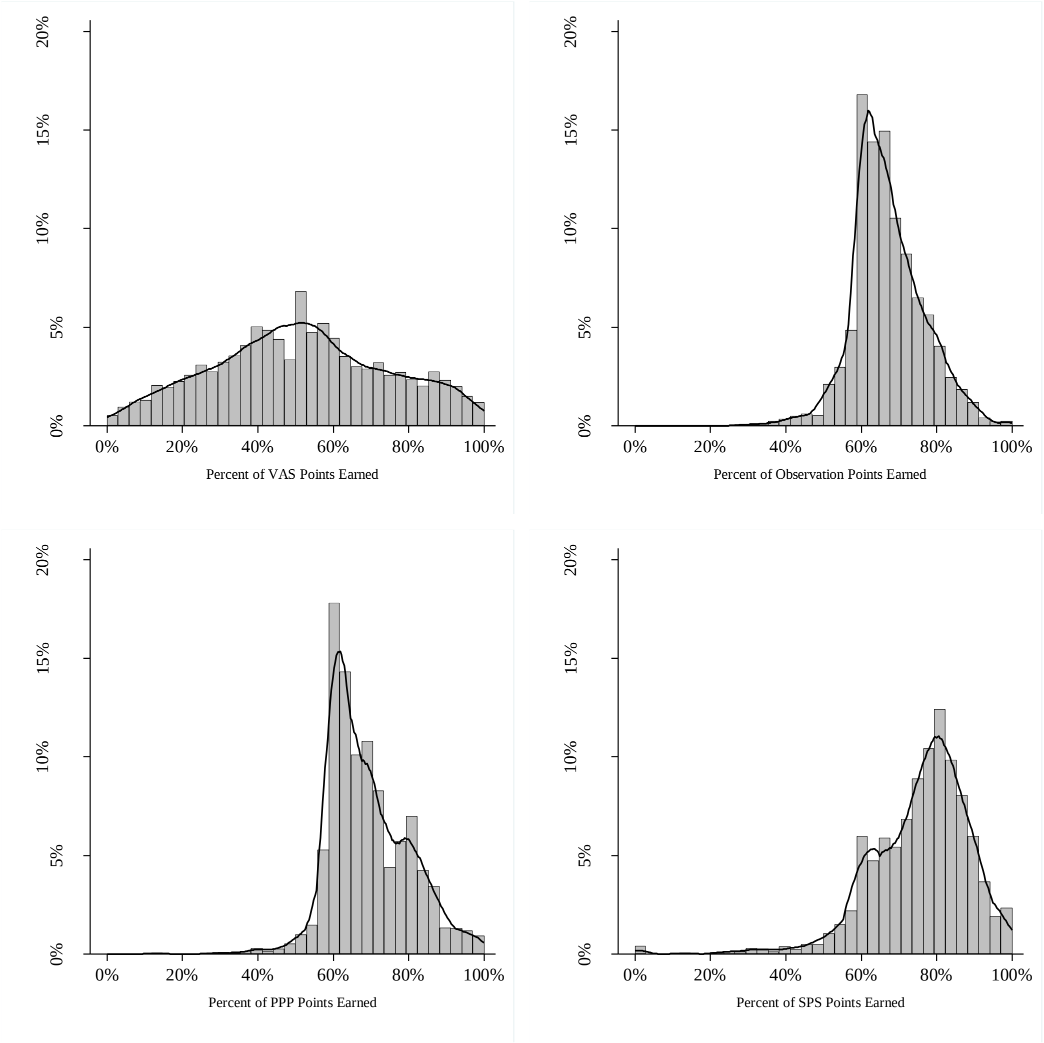


Figure 4. Distributions of percent of VAS, observation, PPP, and SPS points earned for teachers in the Year 1 sample.

Year 2 (2014-2015) Sample

Characteristics of teachers present in the Year 2 sample were similar to those of teachers present in the Year 1 sample. Teachers were predominantly female (71%; *n* = 1,934/2,738), non-URM (56%; *n* = 1,546/2,738), and regular classroom teachers (89%; *n* = 2,434/2,738). Five percent (*n* = 143/2,738) were SE teachers and 6% (*n* = 156/2,738) were bilingual teachers, and the average years of experience was 12.2 (*SD* = 9.59) (see Table 3, again).

The classroom compositions for teachers present in Year 2 were also similar to those of teachers present in the Year 1 sample (see Table 4, again). On average, teachers taught a slightly higher number of classes (*M* = 2.4), though nearly the same number of students across all classes (*M* = 73.1). Approximately half of the students were male (51%), and the majority qualified for FRL (76%) and were URM (76%). Nineteen percent of students were labeled as ELL, 14% were labeled as SE, and 5% were labeled as gifted.

Like the teachers in the Year 1 sample, the distribution of teachers' overall teacher evaluation ratings was normal (i.e., a bell curve), with the majority of teachers earning a score of "Effective" (44.0%; *n* = 1,206/2,738) fewer and similar percent of teachers earning scores of "Minimally Effective" (26.4%; *n* = 722/2,738) and "Highly Effective" (24.0%; *n* = 658/2,738), respectively; and even fewer teachers and similar percent of teachers earning scores of "Ineffective" (2.6%, *n* = 71/2,738) and "Exemplary" (3.0%, *n* = 82/2,738) (see Figure 3, again). In line with teachers' overall ratings, the average summative score was a 3.0 (*SD* = 0.85). On average, teachers in the Year 2 sample earned just over half of the possible VAS points (*M* = 0.53, *SD* = 0.240), 70% of the possible classroom observation points (*M* = 0.70, *SD* = 0.100), just over 70% of the possible PPP points (*M* = 0.73, *SD* = 0.110), and over 80% of the possible SPS points (*M* = 0.82, *SD* = 0.101) (see Table 5, again). The distribution of the percent of VAS points earned was relatively normal, though slightly flatter than in Year 1. The other three measures were more negatively skewed, like in Year 1 (see Figure 5).

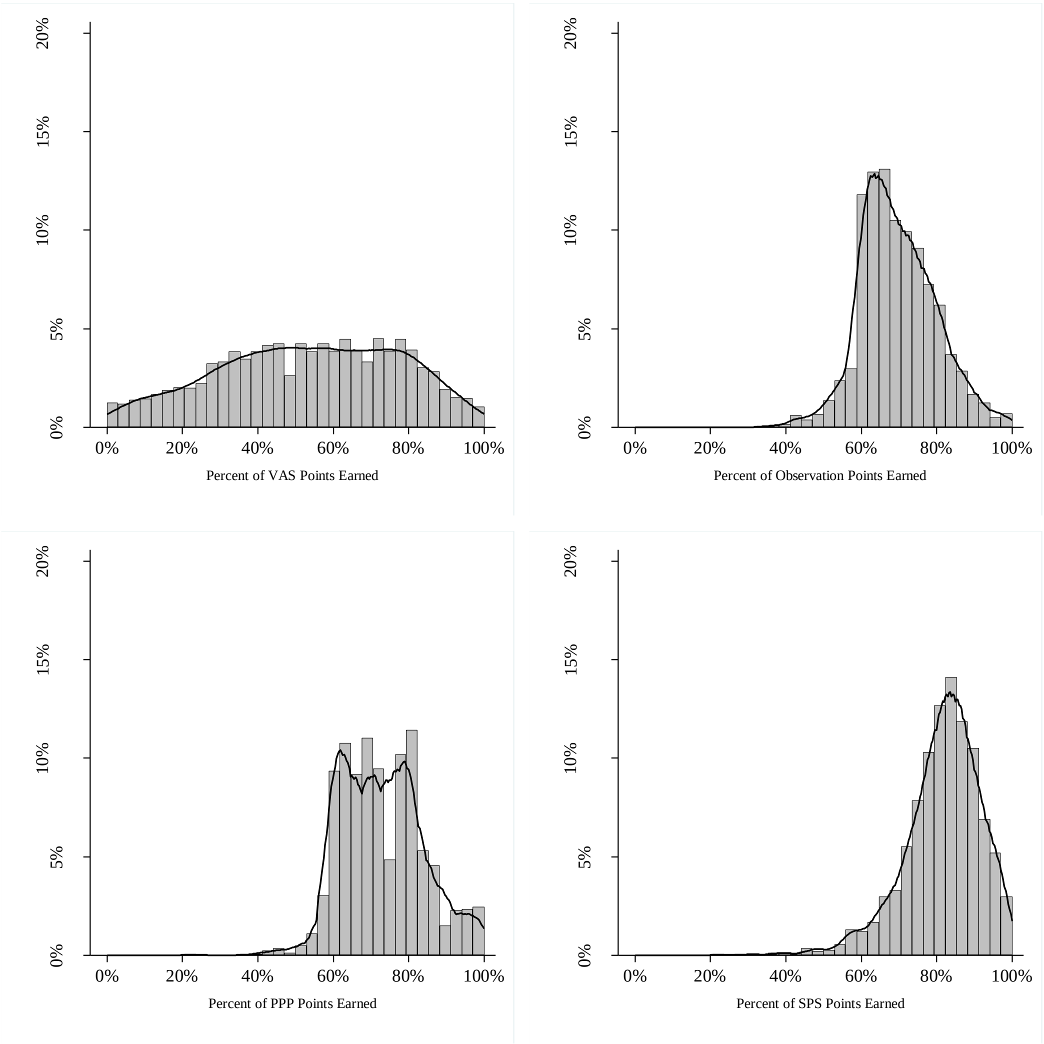


Figure 5. Distributions of percent of VAS, observation, PPP, and SPS points earned for teachers in the Year 2 sample.

Year 3 (2015-2016) Sample

The composition of teachers present in the Year 3 sample differed descriptively from that of teachers in Year 1 and Year 2 on several demographic variables. Just under 60% of teachers were female (59%; *n* = 5,311/8,963), though 22% (*n* = 1,968/8,963) of teachers in Year 3 had missing gender data. Most teachers were non-URM (62%; *n* = 5,546/8,963), and only 6% (*n* = 551/8,963) and 4% (*n* = 341/8,963) were SE teachers and bilingual teachers, respectively. Unlike teachers in Years 1 and 2, the average years of experience for teachers was 10.9 (*SD* = 9.04) (see Table 3, again).

Similar to the teachers present in the Year 1 and Year 2 sample, on average, teachers in the Year 3 sample taught 2.3 classes; however, the average number of students was higher (*M* = 83.4) (see Table 4, again). The average student composition for teachers in the Year 3 sample was very similar to that of teachers in the Year 1 and Year 2 samples, with 51% of students being male, 76% of students qualifying for FRL, and 74% of students being URM. Sixteen percent of students were labeled as ELL, 15% were labeled as SE, and 6% were labeled as gifted.

Like the teachers in both the Year 1 and Year 2 samples, the distribution of teachers' overall teacher evaluation ratings was normal (i.e., a bell curve), with the majority of teachers earning a score of "Effective" (39.7%; *n* = 3,558/8,963) fewer and similar percent of teachers earning scores of "Minimally Effective" (25.7%; *n* = 2,301/8,963) and "Highly Effective" (24.1%; *n* = 2,158/8,963), respectively; and even fewer teachers and similar percent of teachers earning scores of "Ineffective" (6.7%, *n* = 602/8,963) and "Exemplary" (3.8%, *n* = 344/8,963) (see Figure 3, again). In line with teachers' overall ratings, the average summative score was a 2.9 (*SD* = 0.96). On average, teachers in the Year 3 sample earned just over half of the possible VAS points (*M* = 0.53, *SD* = 0.245), just over 70% of the possible classroom observation points (*M* = 0.72, *SD* = 0.099), just over 70% of the possible PPP points (*M* = 0.73, *SD* = 0.114), and over 80%

of the possible SPS points (*M* = 0.81, *SD* = 0.106) (see Table 5, again). Again, the distribution of the percent of VAS points earned was relatively normal, though slightly flatter than in Year 1 again. The other three measures were more negatively skewed, like in Year 1 and Year 2 (see Figure 6).

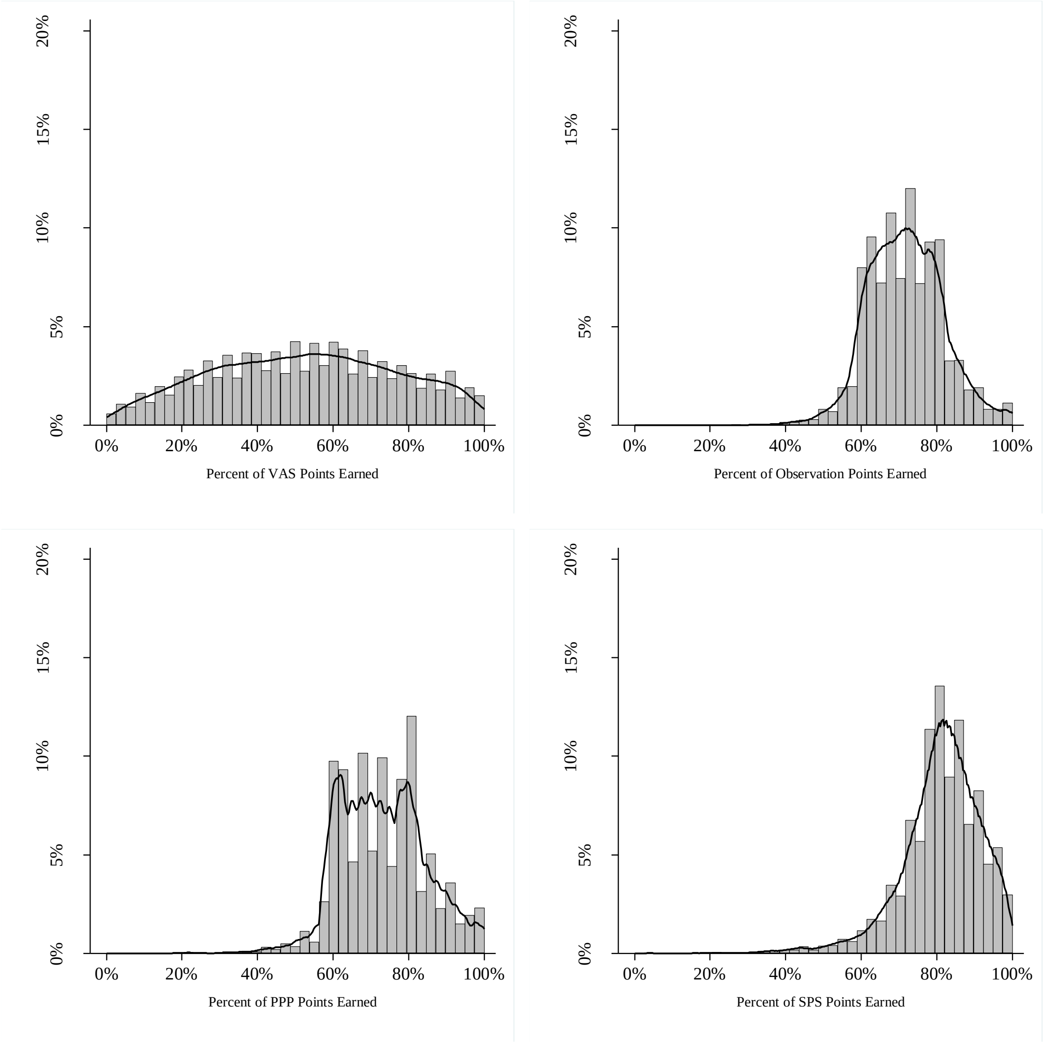


Figure 6. Distributions of percent of VAS, observation, PPP, and SPS points earned for teachers in the Year 3 sample.

Regression Model Results

In this section, I present the results from each of the 48 regression models. This section is grouped into subsections by models per criterion variable. That is, I first discuss all models where the percent of VAS points a teacher earned was the criterion variable, per year and per predictor variable. I then discuss, in turn the classroom observation models, PPP models, and SPS models while following the same structure.

In each subsection, I first examine whether the models overall were significant ($\alpha = .05$). I then discuss whether the main predictor variable in each model (i.e., each of student demographic factors aggregated to the school level) was significantly associated with each model's criterion variable (i.e., each teacher evaluation measure; see Appendix C for full model outputs). Finally, I provide a summary of findings per model, per teacher effectiveness measure. For interpretation purposes, I also reiterate here that all predictor variables that are expressed in terms of percentages (e.g., percent of FRL students, percent of URM students) have been scaled to 10%. That is, when viewing model output and coefficients per model, each coefficient as related to the change in the criterion variable is per a 10% increase/decrease in the predictor variable.

Lastly, it is also worth noting that per model, one or more covariates were significantly associated with the given measure of teacher effectiveness. However, for the purposes of this study, I have limited my results and findings (forthcoming, see Chapters 4 and 5) solely to the student demographic factors aggregated to the school level, as per my research questions. Details about covariate significance, including strength and directionality, are included in Appendix C (see Tables C1-C16).

Model 1 – Percent of VAS Points Earned

Year 1 (2013-2014). Across the four Year 1 models with the percent of VAS points earned as the outcome measures, all were significant ($p < .001$) (see Appendix C, Tables C1-C4). However, the models' fits were all quite poor. As per the R^2 values, only 2.6% to 3.4% of the variance in the percent of teachers' VAS points earned were explained. The below subsections describe the details of each model, given the main predictor variable of interest.

Model 1a – Percent of SE students. The percent of SE students within a teacher's school was significantly and positively associated with the percent of VAS points a teacher earned, when controlling for the school level a teacher taught; a teacher's years of experience; the percent of gifted students within a teacher's classes; the percent of SE students within a teacher's classes; the percent of FRL students both within a teacher's classes and within a school, respectively; the percent of URM students both within a teacher's classes and within a school; and the percent of ELL students within a school. As the percent of SE students within a school increased by 10%, the percent of VAS points a teacher earned increased by 3.3% ($p = .001$).

Model 1b – Percent of ELL students. Similar to the Year 1 SE (1a) model, the percent of ELL students within a teacher's school was also significantly and positively associated with the percent of VAS points a teacher earned, when controlling for the level a teacher taught; a teacher's years of experience; whether a teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school, respectively; the percent of FRL students both within a teacher’s classes and within a school, respectively; and the percent of URM students both within a teacher’s classes and within a school. As the percent of ELL students within a school increased by 10%, the percent of VAS points a teacher earned increased by 1.1% (*p* = .010).

Model 1c – Percent of FRL students. Unlike the Year 1 SE (1a) and ELL (1b) models, the percent of FRL students within a teacher’s school was not significantly associated with the percent of VAS points a teacher earned, when controlling for a teacher’s years of experience; the percent of gifted students within a teacher’s classes; the percent of SE students both within a teacher’s classes and within a school; the percent of FRL students within a teacher’s classes; the percent of ELL students within a school; the percent of URM students within both a teacher’s classes and within a school, respectively; and the number of students within a school. That is, the percent of VAS points a teacher earned did not significantly change based on the percent of FRL students within a teacher’s school.

Model 1d – Percent of URM students. Like the Year 1 FRL (1c) model, the percent of URM students within a school was not significantly associated with the percent of VAS points a teacher earned, when controlling for the level a teacher taught; a teacher’s years of experience; whether the teacher was URM; the percent of gifted students within a teacher’s classes; the percent of SE students both within a teacher’s classes and within a school, respectively; the percent of FRL students both within a teacher’s classes and within a school, respectively; the percent of ELL students within a school; the percent of URM students within a teacher’s classes; and the number of students within a school. That is, the percent of VAS points a teacher earned did not significantly change based on the percent of URM students within a teacher's school.

Year 2 (2014-2015). Across the four Year 2 models, all were significant ($p <$ .001) (see Appendix C, Tables C1-C4). However, the models' fits were all quite poor. As per the R^2 values, only 2.7% to 3.1% of the variance in the percent of teachers' VAS points earned were explained. The below subsections describe the details of each model, given the main predictor variables of interest.

Model 1a – Percent of SE students. The percent of SE students at a school was significantly and positively associated with the percent of VAS points a teacher earned, when controlling for the school level a teacher taught; a teacher's years of experience; the percent of gifted students within a teacher's classes; the percent of SE students within a teacher's classes; the percent of FRL students both within a teacher's classes and within a school, respectively; the percent of URM students both within a teacher's classes and within a school; and the percent of ELL students within a school. As the percent of SE students within a school increased by 10%, the percent of VAS points a teacher earned increased by 3.4% (p = .001).

Model 1b – Percent of ELL students. Similar to the Year 2 SE (1a) model, the percent of ELL students within a school was significantly and positively associated with the percent of VAS points a teacher earned, when controlling for the level a teacher taught; a teacher's years of experience; whether a teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school, respectively; the percent of FRL students both within a teacher’s classes and within a school, respectively; and the percent of URM students both within a teacher’s classes and within a school. As the percent of ELL students within a school increased by 10%, the percent of VAS points a teacher earned increased by 2.2% (*p* < .001).

Model 1c – Percent of FRL students. The percent of FRL students within a school was not significantly associated with the percent of VAS points a teacher earned, when controlling for a teacher’s years of experience; the percent of gifted students within a teacher’s classes; the percent of SE students both within a teacher’s classes and within a school; the percent of FRL students within a teacher’s classes; the percent of ELL students within a school; the percent of URM students within both a teacher’s classes and within a school, respectively; and the number of students within a school. That is, the percent of VAS points a teacher earned did not significantly change based on the percent of FRL students within a teacher’s school.

Model 1d – Percent of URM students. The percent of URM students within a school was significantly and negatively associated with the percent of VAS points a teacher earned, when controlling for the level a teacher taught; a teacher’s years of experience; whether the teacher was URM; the percent of gifted students within a teacher’s classes; the percent of SE students both within a teacher’s classes and within a school, respectively; the percent of FRL students both within a teacher’s classes and within a school, respectively; the percent of ELL students within a school; the percent of URM students within a teacher’s classes; and the number of students within a school. As the percent of URM students within a school increased by 10%, the percent of VAS points a teacher earned decreased by 1.9% (p = .005).

Year 3 (2015-2016). Across the four Year 3 models with VAS as the outcome measure, all were significant (p < .001) (see Appendix C, Tables C1-C4). As per the R^2 values of the Year 3 models, only 3.8% to 4.1% of the variance in the percent of teachers' VAS points earned were explained. The below subsections describe the details of each model, given the main predictor variable of interest.

Model 1a – Percent of SE students. The percent of SE students at a school was significantly and positively associated with the percent of VAS points a teacher earned, when controlling for a teacher's years of experience; the percent of gifted students within a teacher's classes; the percent of SE students within a teacher's classes; the percent of FRL students both within a teacher's classes and within a school, respectively; the percent of URM students both within a teacher's classes and within a school; and the percent of ELL students within a school. As the percent of SE students within a school increased by 10%, the percent of VAS points a teacher earned increased by 1.2% (p = .037).

Model 1b – Percent of ELL students. Unlike the Year 3 SE (1a) model, the percent of ELL students within a school was not significantly associated with the percent of VAS points a teacher earned, when controlling for a teacher's years of experience; whether a teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school, respectively; the percent of FRL students both within a teacher's classes and within a school, respectively; and the percent of URM students both within a teacher's classes and within a school. That is, the percent of VAS points a teacher earned did not significantly change based on the percent of ELL students within a teacher's school.

Model 1c – Percent of FRL students. Similar to the Year 3 SE (1a) model, the percent of FRL students within a school was significantly and positively associated with the percent of VAS points a teacher earned, when controlling for a teacher's years of experience; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school; the percent of FRL students within a teacher's classes; the percent of ELL students within a school; the percent of URM students within both a teacher's classes and within a school, respectively; and the number of students within a school. As the percent of FRL students within a school increased by 10%, the percent of VAS points a teacher earned decreased by 0.8% (p < .001).

Model 1d – Percent of URM students. Similar to the Year 3 ELL (1b) model, the percent of URM students within a school was not significantly associated with the percent of VAS points a teacher earned, when controlling for a teacher's years of experience; whether the teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school, respectively; the percent of FRL students both within a teacher's classes and within a school, respectively; the percent of ELL students within a school; the percent of URM students within a teacher's classes; and the number of students within a school.

That is, the percent of VAS points a teacher earned did not significantly change based on the percent of URM students within a teacher's school.

Model 1 (VAS) summary. Overall, in the majority of models (i.e., n = 7/12; 58%), the student demographic factor was significantly associated with the percent of VAS points a teacher earned (see Table 6). These significant relationships occurred across multiple years, with two in the Year 1 models, three in the Year 2 models, and two in the Year 3 models.

Table 6

Summary of Significance of Main Predictor Variables (PVs) per VAS Model

	Year 1	Year 2	Year 3
Main PV per VAS Model	2013-2014	2014-2015	2015-2016
Percent of SE Students per School	+	+	+
Percent of ELL Students per School	+	+	
Percent of FRL Students per School			-
Percent of URM Students per School		-	

Note: A plus sign (+) in a cell indicates that the main PV in a given year's model was significantly positively associated with the percent of VAS points a teacher earned in that year. A minus sign (-) in a cell indicates that the main PV in a given year's model was significantly negatively associated with the percent of VAS points a teacher earned in that year. A blank cell indicates that the main PV in a given year's model was not significantly associated with the percent of VAS points a teacher earned in that year.

Additionally, each of the four student demographic factors had a significant relationship with the percent of VAS points a teacher earned in at least one of the years, though the directionality of the relationships varied.

Interestingly, when the percent of SE students within a school and the percent of ELL students within a school were significantly associated with the percent of VAS points a teacher earned, both relationships were positive. That is, teachers who taught in schools with higher percentages of SE students and ELL students, respectively, earned higher percentages of VAS points. Although there was only one year where the percent of FRL students within a school and the percent of URM students within in a school, respectively, had significant relationships with the percent of VAS points a teacher earned, these relationships were both negative. The results from these models suggest that, to some degree, the student composition of a school, as per these four demographic factors, affected teachers' VAS scores.

Model 2 – Percent of Classroom Observation Points Earned

Year 1 (2013-2014). Across the four Year 1 models with the percent of classroom observation points as the outcome measures, all were significant ($p < .001$) (see Appendix C, Tables C5-C8). Overall, the models' fits were poor. As per the R^2 values, only 7.2% to 9.2% of the variance in the percent of teachers' observation points earned were explained. The below subsections describe the details of each model, given the main predictor variable of interest.

Model 2a – Percent of SE students. The percent of SE students at a school was not significantly associated with the percent of classroom observation points a teacher earned, when controlling for the school level a teacher taught; a teacher's years of experience; the percent of gifted students within a teacher's classes; the percent of SE students within a teacher's classes; the percent of FRL students both within a teacher's classes and within a school, respectively; the percent of URM students both within a teacher's classes and within a school; and the percent of ELL students within a school. That is, the percent of classroom observation points a teacher earned did not significantly change based on the percent of SE students within a teacher's school.

Model 2b – Percent of ELL students. Unlike the Year 1 SE (2a) model, the percent of ELL students within a school was significantly and negatively associated with the percent of observation points a teacher earned, when controlling for the school level a teacher taught; a teacher's years of experience; whether a teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school, respectively; the percent of FRL students both within a teacher's classes and within a school, respectively; and the percent of URM students both within a teacher's classes and within a school. As the percent of ELL students within a school increased by 10%, the percent of observation points a teacher earned decreased by 0.5% (p = .002).

Model 2c – Percent of FRL students. Like the Year 1 SE (2a) model, the percent of FRL students within a school was not significantly associated with the percent of classroom observation points a teacher earned, when controlling for a teacher's years of experience; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school; the percent of FRL students within a teacher's classes; the percent of ELL students within a school; the percent of URM students within both a teacher's classes and within a school, respectively; and the number of students within a school. That is, the percent of classroom observation points a teacher earned did not significantly change based on the percent of FRL students within a teacher's school.

Model 2d – Percent of URM students. Like the Year 1 SE (2a) and FRL (2c) models, the percent of URM students within a school was not significantly associated with the percent of observation points a teacher earned, when controlling for the school level a teacher taught; a teacher's years of experience; whether the teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school, respectively; the percent of FRL students both within a teacher's classes and within a school, respectively; the percent of ELL students within a school; the percent of URM students within a teacher's classes; and the number of students within a school. That is, the percent of classroom observation points a teacher earned did not significantly change based on the percent of URM students within a teacher's school.

Year 2 (2014-2015). Across the four Year 2 models with the percent of classroom observation points as the outcome measures, all were significant ($p < .001$) (see Appendix C, Tables C5-C8). Overall, the models' fits were weak and similar to those of the Year 1 models. As per the R^2 values, only 7.3% to 8.8% of the variance in the percent of teachers' observation points earned were explained. The below subsections describe the details of each model, per the main predictor variable of interest.

Model 2a – Percent of SE students. The percent of SE students at a school was not significantly associated with the percent of classroom observation points a teacher earned, when controlling for the school level a teacher taught; a teacher's years of experience; the percent of gifted students within a teacher's classes; the percent of SE students within a teacher's classes; the percent of FRL students both within a teacher's classes and within a school, respectively; the percent of URM students both within a teacher's classes and within a school; and the percent of ELL students within a school.

That is, the percent of classroom observation points a teacher earned did not significantly change based on the percent of SE students within a teacher's school.

Model 2b – Percent of ELL students. Similar to the Year 2 SE (2a) model, the percent of ELL students within a school was not significantly associated with the percent of observation points a teacher earned, when controlling for the school level a teacher taught; a teacher's years of experience; whether a teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school, respectively; the percent of FRL students both within a teacher's classes and within a school, respectively; and the percent of URM students both within a teacher's classes and within a school. That is, the percent of classroom observation points a teacher earned did not significantly change based on the percent of ELL students within a teacher's school.

Model 2c – Percent of FRL students. Similar to the Year 2 SE (2a) and ELL (2b) models, the percent of FRL students within a school was not significantly associated with the percent of classroom observation points a teacher earned, when controlling for a teacher's years of experience; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school; the percent of FRL students within a teacher's classes; the percent of ELL students within a school; the percent of URM students within both a teacher's classes and within a school, respectively; and the number of students within a school. That is, the percent of classroom observation points a teacher earned did not significantly change based on the percent of FRL students within a teacher's school.

Model 2d – Percent of URM students. Like all prior Year 2 models (i.e., 2a, 2b, and 2c), the percent of URM students within a school was not significantly associated with the percent of observation points a teacher earned, when controlling for the school level a teacher taught; a teacher's years of experience; whether the teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school, respectively; the percent of FRL students both within a teacher's classes and within a school, respectively; the percent of ELL students within a school; the percent of URM students within a teacher's classes; and the number of students within a school. That is, the percent of classroom observation points a teacher earned did not significantly change based on the percent of URM students within a teacher's school.

Year 3 (2015-2016). Across the four Year 3 models with the percent of classroom observation points as the outcome measures, all were significant ($p < .001$) (see Appendix C, Tables C5-C8). Overall, the models' fits were poor and slightly worse than that of Year 1 and Year 2 models. As per the R^2 values, only 6.2% to 6.3% of the variance in the percent of teachers' observation points earned were explained. The below subsections describe the details of each model, per the main predictor variable of interest.

Model 2a – Percent of SE students. The percent of SE students at a school was not significantly associated with the percent of classroom observation points a teacher earned, when controlling for a teacher's years of experience; the percent of gifted students within a teacher's classes; the percent of SE students within a teacher's classes; the percent of FRL students both within a teacher's classes and within a school, respectively; the percent of URM students both within a teacher's classes and within a school; and the percent of ELL students within a school. That is, the percent of classroom observation points a teacher earned did not significantly change based on the percent of SE students within a teacher's school.

Model 2b – Percent of ELL students. Like the Year 3 SE (2a) model, the percent of ELL students within a school was not significantly associated with the percent of observation points a teacher earned, when controlling for a teacher's years of experience; whether a teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school, respectively; the percent of FRL students both within a teacher's classes and within a school, respectively; and the percent of URM students both within a teacher's classes and within a school. That is, the percent of classroom observation points a teacher earned did not significantly change based on the percent of ELL students within a teacher's school.

Model 2c – Percent of FRL students. Like the Year 3 SE (2a) and ELL (2b) models, the percent of FRL students within a school was not significantly associated with the percent of classroom observation points a teacher earned, when controlling for a teacher's years of experience; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school; the percent of FRL students within a teacher's classes; the percent of ELL students within a school; the percent of URM students within both a teacher's classes and within a school, respectively; and the number of students within a school. That is, the percent of classroom observation points a teacher earned did not significantly change based on the percent of FRL students within a teacher's school.

Model 2d – Percent of URM students. Like the other Year 3 models (i.e., 2a, 2b, and 2c), the percent of URM students within a school was not significantly associated with the percent of observation points a teacher earned, when controlling for a teacher's years of experience; whether the teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school, respectively; the percent of FRL students both within a teacher's classes and within a school, respectively; the percent of ELL students within a school; the percent of URM students within a teacher's classes; and the number of students within a school. That is, the percent of classroom observation points a teacher earned did not significantly change based on the percent of URM students within a teacher's school.

Model 2 (classroom observations) summary. Overall, in the majority of models (i.e., 92%; *n* = 11/12), the student demographic factor was not significantly associated with the percent of observation points a teacher earned (see Table 7). The one significant association occurred in Year 1, when there was a significant and negative relationship between the percent of ELL students in a teacher's school and the percent of classroom observation points a teacher earned.

Table 7

Summary of Significance of Main Predictor Variables (PVs) per Classroom Observation Model

Main PV per Classroom Observation Model	Year 1 2013-2014	Year 2 2014-2015	Year 3 2015-2016
Percent of SE Students per School			
Percent of ELL Students per School	-		
Percent of FRL Students per School			
Percent of URM Students per School			

Note: A plus sign (+) in a cell indicates that the main PV in a given year's model was significantly positively associated with the percent of classroom observation points a teacher earned in that year. A minus sign (-) in a cell indicates that the main PV in a given year's model was significantly negatively associated with the percent of classroom observation points a teacher earned in that year. A blank cell indicates that the main PV in a given year's model was not significantly associated with the percent of classroom observation points a teacher earned in that year.

The results from these models suggest that, overall, the student composition of a school, as per the four student demographic factors of interest in this study, did not affect teachers' classroom observation scores.

Model 3 – Percent of PPP Points Earned

Year 1 (2013-2014). Across the four Year 1 models with the percent of PPP points as the outcome measures, all were significant ($p < .001$) (see Appendix C, Tables C9-C12). Overall, the models' fits were not good. As per the R^2 values, only 9.3% to 10.6% of the variance in the percent of teachers' observation points earned were explained. The below subsections describe the details of each model, given the main predictor variable of interest.

Model 3a – Percent of SE students. The percent of SE students at a school was not significantly associated with the percent of PPP points a teacher earned, when controlling for the school level a teacher taught; a teacher's years of experience; the percent of gifted students within a teacher’s classes; the percent of SE students within a teacher’s classes; the percent of FRL students both within a teacher’s classes and within a school, respectively; the percent of URM students both within a teacher’s classes and within a school; and the percent of ELL students within a school. That is, the percent of PPP points a teacher earned did not significantly change based on the percent of SE students within a teacher’s school.

Model 3b – Percent of ELL students. Similar to the Year 1 SE (3a) model, the percent of ELL students within a school was not significantly associated with the percent of PPP points a teacher earned, when controlling for the school level a teacher taught; a teacher’s years of experience; whether a teacher was URM; the percent of gifted students within a teacher’s classes; the percent of SE students both within a teacher’s classes and within a school, respectively; the percent of FRL students both within a teacher’s classes and within a school, respectively; and the percent of URM students both within a teacher’s classes and within a school. That is, the percent of PPP points a teacher earned did not significantly change based on the percent of ELL students within a teacher’s school.

Model 3c – Percent of FRL students. Similar to the Year 1 SE (3a) and ELL (3b) models, the percent of FRL students within a school was not significantly associated with the percent of PPP points a teacher earned, when controlling for a teacher’s years of experience; the percent of gifted students within a teacher’s classes; the percent of SE students both within a teacher’s classes and within a school; the percent of FRL students within a teacher’s classes; the percent of ELL students within a school; the percent of

URM students within both a teacher’s classes and within a school, respectively; and the number of students within a school. That is, the percent of PPP points a teacher earned did not significantly change based on the percent of FRL students within a teacher’s school.

Model 3d – Percent of URM students. In line with the other Year 1 models (i.e., 3a, 3b, and 3c), the percent of URM students within a school was not significantly associated with the percent of PPP points a teacher earned, when controlling for the school level a teacher taught; a teacher’s years of experience; whether the teacher was URM; the percent of gifted students within a teacher’s classes; the percent of SE students both within a teacher’s classes and within a school, respectively; the percent of FRL students both within a teacher’s classes and within a school, respectively; the percent of ELL students within a school; the percent of URM students within a teacher’s classes; and the number of students within a school. That is, the percent of PPP points a teacher earned did not significantly change based on the percent of URM students within a teacher’s school.

Year 2 (2014-2015).

Model 3a – Percent of SE students. The percent of SE students at a school was not significantly associated with the percent of PPP points a teacher earned, when controlling for the school level a teacher taught; a teacher’s years of experience; the percent of gifted students within a teacher’s classes; the percent of SE students within a teacher’s classes; the percent of FRL students both within a teacher’s classes and within a school, respectively; the percent of URM students both within a teacher’s classes and within a school; and the percent of ELL students within a school. That is, the percent of PPP points a teacher earned did not significantly change based on the percent of SE students within a teacher's school.

Model 3b – Percent of ELL students. Like the Year 2 SE (3a) model, the percent of ELL students within a school was not significantly associated with the percent of PPP points a teacher earned, when controlling for the level a teacher taught; a teacher's years of experience; whether a teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school, respectively; the percent of FRL students both within a teacher's classes and within a school, respectively; and the percent of URM students both within a teacher's classes and within a school. That is, the percent of PPP points a teacher earned did not significantly change based on the percent of ELL students within a teacher's school.

Model 3c – Percent of FRL students. Similar to the Year 2 SE (3a) and ELL (3b) models, the percent of FRL students within a school was not significantly associated with the percent of PPP points a teacher earned, when controlling for a teacher's years of experience; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school; the percent of FRL students within a teacher's classes; the percent of ELL students within a school; the percent of URM students within both a teacher's classes and within a school, respectively; and the number of students within a school. That is, the percent of PPP points a teacher earned did not significantly change based on the percent of FRL students within a teacher's school.

Model 3d – Percent of URM students. Again, in line with the Year 2 SE (3a), ELL (3b), and FRL (3c) models, the percent of URM students within a school was not significantly associated with the percent of PPP points a teacher earned, when controlling for the school level a teacher taught; a teacher's years of experience; whether the teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school, respectively; the percent of FRL students both within a teacher's classes and within a school, respectively; the percent of ELL students within a school; the percent of URM students within a teacher's classes; and the number of students within a school. That is, the percent of PPP points a teacher earned did not significantly change based on the percent of URM students within a teacher's school.

Year 3 (2015-2016).

Model 3a – Percent of SE students. The percent of SE students at a school was significantly and positively associated with the percent of PPP points a teacher earned, when controlling for a teacher's years of experience; the percent of gifted students within a teacher's classes; the percent of SE students within a teacher's classes; the percent of FRL students both within a teacher's classes and within a school, respectively; the percent of URM students both within a teacher's classes and within a school; and the percent of ELL students within a school. As the percent of SE students within a school increased by 10%, the percent of PPP points a teacher earned increased by 1.0% ($p <$ .001).

Model 3b – Percent of ELL students. Unlike the Year 3 SE (3a) model, the percent of ELL students within a school was not significantly associated with the percent of PPP points a teacher earned, when controlling for a teacher's years of experience; whether a teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school, respectively; the percent of FRL students both within a teacher's classes and within a school, respectively; and the percent of URM students both within a teacher's classes and within a school. That is, the percent of PPP points a teacher earned did not significantly change based on the percent of ELL students within a teacher's school.

Model 3c – Percent of FRL students. Similar to the Year 3 ELL (3b) model, the percent of FRL students within a school was not significantly associated with the percent of PPP points a teacher earned, when controlling for a teacher's years of experience; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school; the percent of FRL students within a teacher's classes; the percent of ELL students within a school; the percent of URM students within both a teacher's classes and within a school, respectively; and the number of students within a school. That is, the percent of PPP points a teacher earned did not significantly change based on the percent of FRL students within a teacher's school.

Model 3d – Percent of URM students. Similar to the Year 3 SE (3a) model, the percent of URM students within a school was significantly and negatively associated with the percent of PPP points a teacher earned, when controlling for a teacher's years of experience; whether the teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school, respectively; the percent of FRL students both within a teacher's classes and within a school, respectively; the percent of ELL students within a school; the percent of URM students within a teacher's classes; and the number of students within a school. As the percent of URM students within a school increased by 10%, the percent of PPP points a teacher earned decreased by 0.4% (p = .031).

Model 3 (PPP) summary. Overall, in the majority of models (i.e., 83%; n = 10/12), the student demographic factor was not significantly associated with the percent of PPP points a teacher earned (see Table 8). The two significant associations occurred in Year 3, where there was a significant and positive association between the percent of SE students in a teacher's school and the percent of PPP points a teacher earned and a significant and negative association between the percent of URM students in a teacher's school and the percent of PPP points a teacher earned.

Table 8

Summary of Significance of Main Predictor Variables (PVs) per PPP Model

Main PV per PPP Model	Year 1 2013-2014	Year 2 2014-2015	Year 3 2015-2016
Percent of SE Students per School			+
Percent of ELL Students per School			
Percent of FRL Students per School			
Percent of URM Students per School			-

Note: A plus sign (+) in a cell indicates that the main PV in a given year's model was significantly positively associated with the percent of PPP points a teacher earned in that year. A minus sign (-) in a cell indicates that the main PV in a given year's model was significantly negatively associated with the percent of PPP points a teacher earned in that year. A blank cell indicates that the main PV in a given year's model was not significantly associated with the percent of PPP points a teacher earned in that year.

The results from these models suggest that, overall, the student composition of a school, as per the four student demographic factors of interest in this study, did not affect teachers' PPP scores.

Model 4 – Percent of SPS Points Earned

Year 1 (2013-2014). Across the four Year 1 models with the percent of SPS points as the outcome measures, all were significant ($p < .001$) (see Appendix C, Tables C13-C16). Overall, the models' fits were quite poor. A per the R^2 values, only 1.3% to 3.0% of the variance in the percent of teachers' SPS points earned were explained. The below subsections describe the details of each model, per the main predictor variable of interest.

Model 4a – Percent of SE students. The percent of SE students at a school was significantly and negatively associated with the percent of SPS points a teacher earned, when controlling for the school level a teacher taught; a teacher's years of experience; the percent of gifted students within a teacher's classes; the percent of SE students within a teacher's classes; the percent of FRL students both within a teacher's classes and within a school, respectively; the percent of URM students both within a teacher's classes and within a school; and the percent of ELL students within a school. As the percent of SE students within a school increased by 10%, the percent of SPS points a teacher earned decreased by 1.9% ($p = .002$).

Model 4b – Percent of ELL students. Unlike the Year 1 SE (4a) model, the percent of ELL students within a school was not significantly associated with the percent of SPS points a teacher earned, when controlling for the school level a teacher taught; a teacher’s years of experience; whether a teacher was URM; the percent of gifted students within a teacher’s classes; the percent of SE students both within a teacher’s classes and within a school, respectively; the percent of FRL students both within a teacher’s classes and within a school, respectively; and the percent of URM students both within a teacher’s classes and within a school. That is, the percent of SPS points a teacher earned did not significantly change based on the percent of ELL students within a teacher’s school.

Model 4c – Percent of FRL students. Similar to the Year 1 FRL (4b) model, the percent of FRL students within a school was not significantly associated with the percent of SPS points a teacher earned, when controlling for a teacher’s years of experience; the percent of gifted students within a teacher’s classes; the percent of SE students both within a teacher’s classes and within a school; the percent of FRL students within a teacher’s classes; the percent of ELL students within a school; the percent of URM students within both a teacher’s classes and within a school, respectively; and the number of students within a school. That is, the percent of SPS points a teacher earned did not significantly change based on the percent of FRL students within a teacher’s school.

Model 4d – Percent of URM students. Similar to the Year 1 ELL (4b) and FRL (4c) models, the percent of URM students within a school was not significantly associated with the percent of SPS points a teacher earned, when controlling for the school level a teacher taught; a teacher’s years of experience; whether the teacher was URM; the percent of gifted students within a teacher’s classes; the percent of SE students both within a teacher’s classes and within a school, respectively; the percent of FRL students both within a teacher’s classes and within a school, respectively; the percent of ELL students within a school; the percent of URM students within a teacher’s classes; and the number of students within a school. That is, the percent of SPS points a teacher earned did not significantly change based on the percent of URM students within a teacher’s school.

Year 2 (2014-2015). Across the four Year 2 models with the percent of SPS points as the outcome measures, all were significant ($p < .001$) (see Appendix C, Tables C13-C16). Overall, the models’ fits were somewhat weak. As per the R^2 values, only 8.7% to 17.8% of the variance in the percent of teachers’ SPS points earned were explained. The below subsections describe the details of each model, per the main predictor variable of interest.

Model 4a – Percent of SE students. The percent of SE students at a school was significantly and negatively associated with the percent of SPS points a teacher earned, when controlling for the school level a teacher taught; a teacher’s years of experience; the percent of gifted students within a teacher’s classes; the percent of SE students within a teacher’s classes; the percent of FRL students both within a teacher’s classes and within a school, respectively; the percent of URM students both within a teacher’s classes and within a school; and the percent of ELL students within a school. As the percent of SE students within a school increased by 10%, the percent of SPS points a teacher earned decreased by 1.0% ($p = .009$).

Model 4b – Percent of ELL students. Like the Year 2 SE (4a) model, the percent of ELL students within a school was significantly associated with the percent of SPS points a teacher earned, when controlling for the school level a teacher taught; a teacher's years of experience; whether a teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school, respectively; the percent of FRL students both within a teacher's classes and within a school, respectively; and the percent of URM students both within a teacher's classes and within a school. As the percent of ELL students in a school increased by 10%, the percent of SPS points a teacher earned increased by 0.8% (p < .001).

Model 4c – Percent of FRL students. Unlike the Year 2 SE (4a) and ELL (4b) models, the percent of FRL students within a school was not significantly associated with the percent of SPS points a teacher earned, when controlling for a teacher's years of experience; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school; the percent of FRL students within a teacher's classes; the percent of ELL students within a school; the percent of URM students within both a teacher's classes and within a school, respectively; and the number of students within a school. That is, the percent of SPS points a teacher earned did not significantly change based on the percent of FRL students within a teacher's school.

Model 4d – Percent of URM students. Similar to the Year 2 FRL (4c) model, the percent of URM students within a school was not significantly associated with the percent of SPS points a teacher earned, when controlling for the school level a teacher taught; a teacher's years of experience; whether the teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher’s classes and within a school, respectively; the percent of FRL students both within a teacher’s classes and within a school, respectively; the percent of ELL students within a school; the percent of URM students within a teacher’s classes; and the number of students within a school. That is, the percent of SPS points a teacher earned did not significantly change based on the percent of URM students within a teacher’s school.

Year 3 (2015-2016). Across the four Year 3 models with the percent of SPS points as the outcome measures, all were significant ($p < .001$) (see Appendix C, Tables C13-C16). Overall, the models’ fits were somewhat poor. As per the R^2 values, only 6.8% to 8.7% of the variance in the percent of teachers’ SPS points earned were explained. The below subsections describe the details of each model, per the main predictor variable of interest.

Model 4a – Percent of SE students. The percent of SE students at a school was not significantly associated with the percent of SPS points a teacher earned, when controlling for a teacher’s years of experience; the percent of gifted students within a teacher’s classes; the percent of SE students within a teacher’s classes; the percent of FRL students both within a teacher’s classes and within a school, respectively; the percent of URM students both within a teacher’s classes and within a school; and the percent of ELL students within a school. That is, the percent of SPS points a teacher earned did not significantly change based on the percent of SE students within a teacher’s school.

Model 4b – Percent of ELL students. Unlike the Year 3 SE (4a) model, the percent of ELL students within a school was significantly associated with the percent of

SPS points a teacher earned, when controlling for a teacher's years of experience; whether a teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school, respectively; the percent of FRL students both within a teacher's classes and within a school, respectively; and the percent of URM students both within a teacher's classes and within a school. As the percent of ELL students within a school increased by 10%, the percent of SPS points a teacher earned increased by 1.9%. ($p < .001$).

Model 4c – Percent of FRL students. Like the Year 3 SE (4a) model, the percent of FRL students within a school was not significantly associated with the percent of SPS points a teacher earned, when controlling for a teacher's years of experience; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school; the percent of FRL students within a teacher's classes; the percent of ELL students within a school; the percent of URM students within both a teacher's classes and within a school, respectively; and the number of students within a school. That is, the percent of SPS points a teacher earned did not significantly change based on the percent of FRL students within a teacher's school.

Model 4d – Percent of URM students. Similar to the Year 3 ELL (4b) model, the percent of URM students within a school was significantly associated with the percent of SPS points a teacher earned, when controlling for a teacher's years of experience; whether the teacher was URM; the percent of gifted students within a teacher's classes; the percent of SE students both within a teacher's classes and within a school, respectively; the percent of FRL students both within a teacher's classes and within a school, respectively; the percent of ELL students within a school; the percent of URM students within a teacher's classes; and the number of students within a school. As the percent of URM students within a school increased by 10%, the percent of SPS points a teacher earned decreased by 0.6% (*p* < .001).

Model 4 (SPS) summary. Overall, in the majority of models (i.e., 58%; *n* = 7/12), the student demographic factor was not significantly associated with the percent of SPS points a teacher earned. However, there were significantly relationships in a number of models, and these relationships occurred over multiple years with one in Year 1, two in Year 2, and two in Year 3 (see Table 9).

Table 9

Summary of Significance of Main Predictor Variables (PVs) per SPS Model

	Year 1	Year 2	Year 3
Main PV per SPS Model	2013-2014	2014-2015	2015-2016
Percent of SE Students per School	-	-	
Percent of ELL Students per School		+	+
Percent of FRL Students per School			
Percent of URM Students per School			-

Note: A plus sign (+) in a cell indicates that the main PV in a given year's model was significantly positively associated with the percent of SPS points a teacher earned in that year. A minus sign (-) in a cell indicates that the main PV in a given year's model was significantly negatively associated with the percent of SPS points a teacher earned in that year. A blank cell indicates that the main PV in a given year's model was not significantly associated with the percent of SPS points a teacher earned in that year.

Additionally, all but one student demographic factors (i.e., the percent of FRL students within a teacher's school) had a significant relationship with the percent of SPS points a teacher earned in at least one of the years, though the directionality of the relationships varied.

Interestingly, when the percent of ELL students within a school was significantly associated with the percent of SPS points a teacher earned, the relationship was positive. That is, teachers who taught in schools with higher percentages of ELL students earned higher percentages of SPS points. Overall, the results from these models suggest that, to some degree, the student composition of a school, as per the percentages of SE, ELL, and URM students, respectively, affected teachers' SPS scores.

Results Summary

In this chapter, I provided results from each of the 48 regression models that determined what student demographic factors aggregated to the school level were significantly associated with the percent of VAS, observation, PPP, and SPS points a teacher earned, respectively. I also provided overall summaries of the results per each teacher evaluation measure. In the next chapter, I present the findings stemming from these results.

CHAPTER 5

FINDINGS

In this chapter, I discuss the findings from this study. I first discuss findings per each of the four measures of teacher effectiveness, and then the findings taken together from a broader perspective. I also offer insights into the presented findings, as well as situate them within the current literature. Lastly, I close this chapter with two possible interpretations of this study's findings.

VAS Measure

Compared to the three other measures, teachers' VAS scores appeared to be the most susceptible to bias by the four student demographic factors aggregated to the school level of interest in this study, based on the numbers of models where school-level student demographic factors were significantly associated with the percent of VAS scores a teacher earned (see Table 10).

Table 10

Comparison of Significance of Main Predictor Variables (PV), per PV, Teacher Evaluation Measure, and Year

Teacher Evaluation Measure	Percent of SE Students	Percent of ELL Students	Percent of FRL Students	Percent of URM Students
Pct. of VAS Points Earned				
Year 1	+	+		
Year 2	+	+		-
Year 3	+		-	
Pct. of Classroom Observation Points Earned				
Year 1				
Year 2		-		
Year 3				
Pct. of PPP Points Earned				
Year 1				
Year 2				
Year 3	+			-
Pct. of SPS Points Earned				
Year 1	-			
Year 2	-	+		
Year 3		+		-

Note: A plus sign (+) in a cell indicates that the main PV in a given year's model was significantly positively associated with the corresponding teacher evaluation measure and year. A minus sign (-) in a cell indicates that the main PV in a given year's model was significantly negatively associated with the corresponding teacher evaluation measure and year. A blank cell indicates that the main PV in a given year's model was not significantly associated with the corresponding teacher evaluation measure and year.

As noted in the prior chapter, the directionality and strength of these associations varied depending on the exact model and year, but all three years of VAS scores were significantly associated with at least one of the student demographic factors.

Overall, the results of these multiple significant associations were not surprising, given Amrein-Beardsley and Geiger's (revise and resubmit) findings of evidence of school-level bias of teacher's VAS scores. However, what was surprising was the number of significant and *positive* associations between teachers' VAS scores and student demographic factors (see Table 11).

Table 11

Breakdown of Predictor Variable (PV) Significance Across All VAS Models

	Total Models	Positively Sig.	Negatively Sig.	No Sig.
School Variables	12 (100%)	5 (42%)	2 (17%)	5 (42%)
Pct. of SE Students	3 (100%)	3 (100%)	0 (0%)	0 (0%)
Pct. of ELL Students	3 (100%)	2 (67%)	0 (0%)	1 (33%)
Pct. of FRL Students	3 (100%)	0 (0%)	1 (33%)	2 (67%)
Pct. of URM Students	3 (100%)	0 (00%)	1 (33%)	2 (67%)

Note: The "Total Models" value indicates the number of VAS models that included each PV. The "Positive Sig.," "Negative Sig.," and "No Sig." columns indicate in how many models the specified PV was either significantly and positively associated, significantly and negatively associated, or not significantly associated with the percent of VAS points a teacher earned. Percentages in parentheses represent the proportion of models where a given PV was significantly and positively associated, significantly and negatively associated, or not significantly associated with the percent of VAS points a teacher earned. All counts are across all three years of models. Percentages might not sum to 100 due to rounding.

Out of the demographic factors across all models, the percent of SE students within a school was always significantly and positively associated with the percent of VAS points a teacher earned. A similar pattern was found for the percent of ELL students within a school; however, that factor was only significant in two of the three models.

Both of these results are curious, given that the majority of prior research has found a negative association between SE and ELL students and teachers' VAM scores, respectively (e.g., Amrein-Beardsley & Geiger, revise and resubmit; Ballou & Springer, 2015; Newton et al., 2010). One simple possible explanation was that between during the 2013-2014 through 2015-2016 school years, SE and ELL students, respectively, performed just as well or better than non-SE and non-ELL students, respectively, on the standardized tests used in teachers' VAS score calculations. However, I found this unlikely to be the case. Per data from the NMPED (2019a), in the 2015-2016 school year, 16% of SE and ELL students combined were proficient or better on reading assessments compared to 45% of non-SE/non-ELL students. I noted the same pattern for mathematics assessments, as 7% of SE and ELL students combined were proficient or better compared to 25% of non-SE/non-ELL students. While these percentages are solely descriptive, it provides enough evidence to counter the possibility that higher SE and ELL student test scores could explain these surprising regression findings.

Another possible explanation for this finding, at least in comparison to Amrein-Beardsley and Geiger's work (revise and resubmit), stems from the inclusion of control variables. In Amrein-Beardsley and Geiger's prior work, no potential confounding variables were considered when they found that teachers in schools with relatively higher proportions of SE and ELL students, respectively, had significantly lower VAS scores than teachers in schools with relatively lower proportions of these students. That is, their finding resulted from analyses that did not take any additional factors that could affect a teacher's VAS estimate into consideration (e.g., the number of SE or ELL students a teacher taught across his or her classes, school size vis-à-vis student enrollment). It is possible that controlling for a variety of potentially confounding variables, as I did in this study, resulted in such different statistical estimates that the directionalities of the significant associations were reversed.

Regarding this difference in directionalities between studies, it is difficult, if not impossible to say whose result is “correct,” or, perhaps, more accurate, given that different methods were used in both studies. On one hand, using multiple regression should produce less biased estimates than *t*-tests or fixed effects ANOVA (i.e., what Amrein-Beardsley and Geiger used) as multiple regression allows for holding other (potentially confounding) variables constant, thereby reducing the amount of unexplained variance in the outcome variable (e.g., the percent of a teacher’s VAS points earned). However, the likelihood that I did not account for other factors that affected the percent of VAS points a teacher earned was high, given the models’ very low R^2 values, the limited information in the NMPED datasets, and the lack of transparency into the actual model used to generate teachers’ VAS scores. Notwithstanding the importance of the directionality of a significant association between a teacher’s VAS score and any confounding factor, especially given how this was of major concern in the *State ex rel. Stewart v. New Mexico Public Education Department* (2015) lawsuit, I argue that the most important aspect of this finding is the mere presence of any significant association between any demographic factor and teachers’ VAS scores.

As previously discussed, VAMs are likely to have issues with both validity *and* reliability, which can subsequently affect the likelihood of a biased estimate being generated for any given teacher. While the directionality of a significant association certainly affects the validity of a VAM estimate, the inconsistent (i.e., unreliable) nature of a VAM estimate makes the question of directionality less important. That is, practically speaking, if certain teachers' VAM scores are significantly associated with a construct-irrelevant factor (e.g., a student demographic factor), these teachers are not fairly evaluated regardless of the directionality of the association. Further, given the statistical nature of a normal distribution, if one teacher's VAM estimate is biased, all teachers' VAM estimates are biased due each estimate being derived relative to others. Therefore, these results serve to underscore VAMs' overall unreliable nature and high levels of susceptibility to bias more so than anything else.

Classroom Observation and PPP Measures

Compared to the other measures of teacher effectiveness, teachers' classroom observation and PPP scores seemed to be the ones least likely to be biased by student demographic factors aggregated to the school level (see Table 10, again). Additionally, this conclusion becomes more apparent when examining the number of significant predictors across all observation and PPP models, respectively (see Tables 12-13).

Table 12

Breakdown of Predictor Variable (PV) Significance Across All Classroom Observation Models

	Total Models	Positively Sig.	Negatively Sig.	No Sig.
School Variables	12 (100%)	0 (0%)	1 (8%)	11 (92%)
Pct. of SE Students	3 (100%)	0 (0%)	0 (0%)	3 (100%)
Pct. of ELL Students	3 (100%)	0 (0%)	1 (33%)	2 (67%)
Pct. of FRL Students	3 (100%)	0 (0%)	0 (0%)	3 (100%)
Pct. of URM Students	3 (100%)	0 (0%)	0 (0%)	3 (100%)

Note: The "Total Models" value indicates the number of classroom observation models that included each PV. The "Positive Sig.," "Negative Sig.," and "No Sig." columns indicate in how many models the specified PV was either significantly and positively associated, significantly and negatively associated, or not significantly associated with the percent of classroom observation points a teacher earned. Percentages in parentheses represent the proportion of models where a given PV was significantly and positively associated, significantly and negatively associated, or not significantly associated with the percent of classroom observation points a teacher earned. All counts are across all three years of models. Percentages might not sum to 100 due to rounding.

Table 13

Breakdown of Predictor Variable (PV) Significance Across All PPP Models

	Total Models	Positively Sig.	Negatively Sig.	No Sig.
School Variables	12 (100%)	1 (8%)	1 (8%)	10 (83%)
Pct. of SE Students	3 (100%)	1 (33%)	0 (0%)	2 (67%)
Pct. of ELL Students	3 (100%)	0 (0%)	0 (0%)	3 (100%)
Pct. of FRL Students	3 (100%)	0 (0%)	0 (0%)	3 (100%)
Pct. of URM Students	3 (100%)	0 (0%)	1 (33%)	2 (67%)

Note: The "Total Models" value indicates the number of PPP models that included each PV. The "Positive Sig.," "Negative Sig.," and "No Sig." columns indicate in how many models the specified PV was either significantly and positively associated, significantly and negatively associated, or not significantly associated with the percent of PPP points a teacher earned. Percentages in parentheses represent the proportion of models where a given PV was significantly and positively associated, significantly and negatively associated, or not significantly associated with the percent of PPP points a teacher earned. All counts are across all three years of models. Percentages might not sum to 100 due to rounding.

The regression results, in combination with the fact that classroom observations are mostly based on a teacher's behaviors, attitudes, practices, and interactions that occur within the confines of his or her classroom(s), serve to substantiate the claim that the various student demographic factors at the school level of interest in this study were unlikely to bias teachers' observation or PPP scores (see Tables 12-13).

The one significant and positive association across all of the observation and PPP models—the percent of SE students within a teacher's school and the percent of PPP points earned—is possibly of note, given its directionality. This directionality is puzzling, also as similar to the significant and positive relationships between this variable and the percent of VAS points teachers earned. Given this pattern, it might be possible that teachers who teach in schools with higher proportions of SE students are more likely to have their scores upwardly biased, albeit for currently unknown reasons. Or, another possibility is that teachers who teach in schools with higher proportions of SE students truly are more effective than teachers who teach in schools with lower proportions of SE students.

Worthier of discussion, however, is the overall lack of significant associations between the school-level student demographic factors and the percent of observation and PPP points a teacher earned, respectively. Similar to some of the VAS score results, these also are contrary to Amrein-Beardsley and Geiger's (revise and resubmit) prior findings that indicated that teachers in schools with relatively higher proportions of SE, ELL, FRL, and URM students, respectively, had significantly higher observation and PPP scores than teachers in schools with relatively lower populations of these students. However, again, they did not control for potentially confounding factors, so it is possible that their significant differences were due to Type I errors (i.e., false positives) rather than being true indicators that teachers' observation and PPP scores actually differed based on the examined student demographic factors at the school level.

The overall lack of significant associations is also curious given it counters prior research (e.g., Blazar et al., 2016; Campbell & Ronfeldt, 2018), as well. One possibility that might, at least partially, further explain this lack is that one group of important factors is missing from the analyses: the characteristics of the observers (see Bailey, Bocala, Shakman, & Zweig, 2016). Not only are these factors missing from this study, but they are frequently missing from other research and discussions about teacher effectiveness as related to observation and PPP scores.

While I controlled for some of the characteristics of the students across a teacher's classes, along with teacher-specific demographic variables, I was unable to control for any observer characteristics (e.g., race/ethnicity, gender, years of teaching experience, etc.). There are likely complex and intertwined implicit assumptions, stereotypes, and biases at play among an observer's characteristics, the teacher's characteristics, and the characteristics of the students in a classroom, of which separation is impossible (Nasir & Hand, 2006). Even when controlling for several student and teacher characteristics, the lack of observer characteristics could potentially account for some of the unexplained variance, and lack of significance, between the four student demographic factors of interest in this study and teachers' observation and PPP scores.

Given the numerous ways that classroom observations can be biased by the observer (i.e., rater bias; Hoyt, 2000), it would not be surprising that different observers rate teachers differently based on the teacher's and students' (demographic) characteristics, along with the (likely) implicit stereotypes and judgments that an observer has about those teachers and students (e.g., Gershenson, Holt, & Papageorge, 2015; Jordan-Irvine, 1990; Peterson, Rubie-Davies, & Osborne, & Sibley, 2016; Van den Bergh, Denessen, Hornstra, Voeten, & Holland, 2010; see also Nasir & Hand, 2006). An observer might unknowingly place different value judgments on witnessed student behavior or witnessed interactions between a teacher and a student, based on the observer's (likely subconscious) beliefs about how different types of students or teachers should behave, act, emote, and the like. These subconscious beliefs are often, at least in part, affected or shaped by the observer's own identities.

For example, white teachers—the majority of whom make up the teaching workforce (Snyder, de Brey, & Dillow, 2019)—have frequently rated African American and Latinx students as more disruptive, less attentive, and less intelligent than white students (e.g., Bates & Glick, 2013; McGrady & Reynolds, 2013; Wodtke, 2012; Wright, 2016), even when non-white and white students have exhibited the same or very similar types of behavior (Okonofua & Eberhardt, 2015). While African American and other non-white teachers hold their own different stereotypes and implicit assumptions about students of varying races and/or ethnicities, compared to white teachers, African American teachers are less likely to rate African American students as unintelligent, lazy, or disruptive compared to white students (Quinn & Stewart, 2019). Teachers who hold implicit stereotypes about different types of students are often likely to communicate those value judgments—whether they are positive or negative—through their body language, tones of voice, and other subtle and nuanced manners (Babad, 1993; Babad, Bernieri, & Rosenthal, 1991). These judgments, which are detectable by students, can subsequently affect students' behaviors, interactions, and attitudes in the classroom (Babad et al., 1991; Dovidio, Kawakami, & Gaertner, 2002; McKown & Weinstein, 2003; Wheeler & Petty, 2001). By extension, it is not unreasonable to think that observers would hold similar implicit stereotypes and attitudes as described above, as such beliefs are often unconscious, unintentional, and widespread (see Greenwald & Banaji, 1995). As such, a teacher's observation or PPP scores could very well be affected by an observer’s unconscious bias (e.g., cultural bias, ethnic bias; Chang & Sue, 2003; see also Bailey et al., 2016; McGrady & Reynolds, 2013).

Teachers who teach higher proportions of FRL or URM students are the most likely to be susceptible to such bias, as white middle-class standards of behavior, attitudes, and achievement are what is commonly accepted as the norm (Carter, 2003; Morris, 2005; Villegas, 1988), even by teachers and administrators who are not white or middle class. Therefore, it seems quite plausible that the percent of observation and PPP points a teacher earned might have been affected by observer characteristics, or the interaction among observer, teacher, and student characteristics. This line of inquiry should be pursued further, especially in states, districts, or schools where there are higher proportions of students, teachers, and observers who differ in racial/ethnic identity, among others (e.g., gender, class, ability).

In addition to the above, further complicating teachers’ observation scores, which might have also played a part in the lack of significant findings, include the relative infrequency of observations (Herlihy et al., 2014), which can add to the subjective nature of scores (e.g., was a teacher observed on a day where students were more or less motivated, more or less engaged, and the like); the likelihood of the purposeful manipulation of scores, either due to purported benevolent reasons (e.g. artificially inflating a teacher’s score to ensure the teacher received formative feedback instead of a punitive developmental plan; Kraft & Gilmour, 2016) or more suspect reasons (e.g., artificially inflating, deflating[7], or conflating scores due to pressure from high stakes accountability policies; Amrein-Beardsley & Geiger, 2019b; see also Campbell, 1976); observers' differences in opinions about the main purpose of classroom observations (i.e., developmental or summative; Bell et al., 2018; Gabriel, 2018; Gabriel & Woulfin, 2017); teachers' expertise and nuanced pedagogical practices being reduced to rubric-driven actions (e.g., Amrein-Beardsley, Holloway-Libell, Cirell, Hays, & Chapman, 2015); and, among many others, the concentrations of less effective teachers in classrooms and schools with higher proportions of SE, ELL, FRL, and URM students (Borman & Kimball, 2005; Goldhaber, Lavery, & Theobald, 2015; Goldhaber et al., 2018; Kalogrides, Loeb, & Beteille, 2013). As per this study, none of these additional possible cofounding factors was taken into consideration, mostly due to the data not being captured or known.

In sum, there are a multitude of factors that can bias teachers' collective classroom observation and PPP scores, including those discussed herein, and others (e.g., class subject). While many teachers and administrators prefer classroom observations to VAMs, for a variety of reasons (e.g., Collins, 2014; Goldring et al., 2015), the likelihood of rater bias is incredibly high, even when observers have been trained and/or certified, have high interrater reliability, and are following clearly defined rubrics. While such bias is often not purposeful, it can affect nearly every aspect of every observation of every teacher in every school. There is an entire body of literature that examines the complex

[7] Per Amrein-Beardsley and Geiger (2019b), artificial inflation is defined as "any source of manipulation that causes a spurious increase in an indicator" (p. 472). Artificial deflation is defined as "any source of external manipulation that causes a spurious decrease in a social indicator that is not due to a 'true' or authentic decline," or the "inverse of artificial inflation" (p. 474).

interplay among teacher, student, and observer identities; implicit stereotypes and assumptions based on those identities; and differential treatment of students by teachers based on these stereotypes, assumptions, and identities. While the findings from this study indicate that teachers' observation and PPP scores are unlikely to be biased by the student demographic factors of interest in this study, I caution that these results only be interpreted within the very specific confines of this study as the likelihood of bias across these measures is high.

SPS Measure

Compared to the three other measures, teachers' SPS scores appeared to be the second most susceptible to bias by the four student demographic factors aggregated to the school level of interest in this study, based on the numbers of models where these factors were significantly associated with the percent of SPS scores a teacher earned (see Table 10, again). Similar to the VAS models, the directionality of the significant associations varied depending on the exact model and year.

Across the five models where a student demographic factor had a significant relationship with the percent of SPS points a teacher earned, 60% (i.e., *n* = 3/5) were significant and negative and while the remaining 40% (i.e., *n* = 2/5) were significant and positive (see Table 14).

Table 14

Breakdown of Predictor Variable (PV) Significance Across All SPS Models

	Total Models	Positively Sig.	Negatively Sig.	No Sig.
School Variables	12 (100%)	2 (17%)	3 (25%)	7 (58%)
Pct. of SE Students	3 (100%)	0 (0%)	2 (67%)	1 (33%)
Pct. of ELL Students	3 (100%)	2 (67%)	0 (0%)	1 (33%)
Pct. of FRL Students	3 (100%)	0 (0%)	0 (0%)	3 (100%)
Pct. of URM Students	3 (100%)	0 (0%)	1 (33%)	2 (67%)

Note: The "Total Models" value indicates the number of SPS models that included each PV. The "Positive Sig.," "Negative Sig.," and "No Sig." columns indicate in how many models the specified PV was either significantly and positively associated, significantly and negatively associated, or not significantly associated with the percent of SPS points a teacher earned. Percentages in parentheses represent the proportion of models where a given PV was significantly and positively associated, significantly and negatively associated, or not significantly associated with the percent of SPS points a teacher earned. All counts are across all three years of models. Percentages might not sum to 100 due to rounding.

Two of the three significant and negative associations were between the percent of SE students within a school and the percent of SPS points a teacher earned (with the third being the percent of URM students within a school), and all of the significant and positive associations were between the percent of ELL students within a school and percent of SPS points a teacher earned.

Interestingly, the significant and negative relationship between the percent of SE students within a school and the percent of SPS points a teacher earned was counter to what Amrein-Beardsley and Geiger (revise and resubmit) found, as in their study, teachers who taught in schools with a relatively higher percent of SE students had significantly higher SPS scores, although this significant difference was only noted for

Year 3. Also of note here are the lack of significant associations between the percent of FRL students and teachers’ SPS scores, as previously, teachers who taught in schools with higher relatively proportions of FRL students had significantly higher SPS scores across all three years (Amrein-Beardsley & Geiger, revise and resubmit).

The moderate likelihood of teachers’ SPS scores being biased by student demographic factors aggregated to the school level is somewhat puzzling, given that student surveys function in a way similar to that of classroom observations (i.e., SPS scores are likely informed mostly from interactions within the confines of a teacher’s classroom(s)). These potentially peculiar effects could be due to a combination of measurement and related issues.

First and foremost, there is somewhat limited evidence that SPSs, including that which was used in New Mexico (i.e., a form of the Tripod), have been psychometrically validated and externally vetted (Geiger & Amrein-Beardsley, 2019). While the Tripod developer has stated that the instrument is “research-based” (Ferguson & Danielson, 2014, p. 101), there is no readily apparent evidence, if any at all, that supports this claim. While there is little doubt that the Tripod was thoughtfully developed and created, it is difficult to assess to what extent measurement concerns are in play without knowing exactly how and to what extent the instrument was psychometrically validated (see Geiger & Amrein-Beardsley, 2019).

Since the Tripod was created in 2001, researchers have conducted multiple studies to examine its reliability and validity. However, these studies all used the same dataset (i.e., that from the MET study), which means that any major (or minor) flaws or idiosyncrasies related to that data would have possibility permeated all subsequent studies that drew upon the same data source. Additionally, typically the only evidence of SPS validity that was assessed was convergent-related evidence of validity, as researcher most often examined the relationships between teachers' Tripod scores and other measures of teacher effectiveness or student achievement (Kane & Cantrell, 2010; Kane & Staiger, 2012; Kuhfeld, 2017; Raudenbush & Jean, 2014; Sandilos et al., 2019; Wallace et al., 2016). The few studies where researchers have examined Tripod data in isolation resulted in somewhat differing findings regarding, for example, the underlying factor structure of the survey (Ferguson, 2010; Ferguson & Danielson, 2014; Schweig, 2014; Wallace et al., 2016).

The relative dearth of general SPS and Tripod-specific information, especially regarding performance across different types of students, teachers, and schools, make the interpretation of this study's results somewhat difficult. For example, it is unknown if the Tripod survey underwent differential item functioning (DIF) analyses (i.e., whether specific survey items function differently based student characteristics that are unrelated to the construct being measured; see AERA et al., 2014), which could potentially explain some of the significant associations, or lack thereof, between student demographic factors and teachers' SPS scores.

Compounding the above concerns could also be the likelihood of teachers' SPS scores being influenced by the gender and racial/ethnic identities of both the students and the teacher in a given class (Basow et al., 2006; Smith et al., 2007). Similar to classroom observation measures, unconscious biases and implicit stereotypes could be in play with

SPSs as well. There have been noted differences in students' ratings of teachers based on the demographic factors of both groups, as well as differences in ratings based on students' perceptions of teachers' attributes that are in no way related to teachers' pedagogical effectiveness or socioemotional qualities related to teaching (e.g., a teacher's perceived attractiveness, sexual orientation, manner of dress; Gurung & Vespia, 2007; Hamermesch & Parker, 2005). While the majority of this research has been conducted in higher education settings, it is unreasonable to expect that elementary and secondary education settings are immune to the numerous potential confounding factors.

Additionally, a lot of the same effects that can bias the observers who conduct classroom observations can also bias students who take the surveys (e.g., halo effect, fatal flaw effect). These effects might even be magnified in the context of SPSs, since students are not trained to rate their teachers in the same way classroom observers are (Wallace et al., 2016). And as previously noted, another complication might also be the extent to which children of different ages are qualified and capable to rate their teachers (De Jong & Westerhof, 2001; Fauth et al., 2014; Kuhfeld, 2017; Kunter & Baumert, 2006; Liaw & Goh, 2003), which, in the case of New Mexico, also informs whether a parent will complete the survey on a student's behalf. This different subset of survey takers could have resulted in additional noise as I was unable to control for specific grade level (and therefore whether the student or his or her parent took the survey).

All things considered, it is not of great surprise that there were a handful of significant associations between the student demographic factors at the school level of interest in this study and the percent of SPS points a teacher earned. However, and especially compared to the other measures of teacher effectiveness, more research into the reliability, validity, and potentially biased nature of SPSs within their current settings (e.g., schools of different student body compositions, grade levels, subjects) is desperately needed. While surveys in general, along with their related elements (e.g., design, psychometric assessment, response rates, response bias, and the like), have been widely studied, they have only very rarely been examined in the context of elementary and secondary school students evaluating teachers. While SPSs undoubtedly provide unique and likely formative feedback to teachers, until they are validated specifically for their current uses—and especially if they are to be used in potentially high stakes settings or for summative purposes (see AERA et al., 2014)—their results should be interpreted with utmost caution.

Summary of Findings

The overarching finding from this study is that the teacher evaluation measures used in New Mexico between the 2013-2014 and 2015-2016 school years had multiple significant relationships with a variety of school-level student demographic factors of interest in this study. These relationships varied in directionality, strength, and consistency based on the teacher effectiveness measure, the specific demographic factor, and the school year. Teachers' VAS scores had the highest number of significant relationships with student demographic factors at the school level, followed by SPS scores, PPP scores, and classroom observation scores.

As such, I draw two main interpretations from these findings. The first interpretation is that New Mexico teachers' VAS, observation, PPP, and SPS scores between the 2013-2014 and 2015-2016 school years evidenced statistical bias due to a variety of student demographic factors at the school level, though this bias occurred in varying degrees and directions, and for potentially varying reasons. The second interpretation is that the significant relationships between the school-level student demographic factors and measures of teacher effectiveness were due to actual differences in teacher quality. That is, teachers' levels of effectiveness truly differed based on the type(s) of students within their schools. The following subsections discuss each of these interpretation in turn, along with possible implications of each.

Interpretation 1: Biased measures. Among the 48 models, although the main predictor variable under analysis was not always statistically significantly associated with the respective teacher effectiveness measure variable, each model did in fact have at least one variable that was significantly associated with the teacher effectiveness measure (see Appendix C, Tables C1-C16). While coefficients of some of the significant factors may seem small and therefore of minimal pragmatic importance, the presence of any significant relationships is a signal of a bigger problem. All of the predictor variables and included covariates should be "irrelevant" to the teaching effectiveness construct, and therefore these variables should not "…differentially affect the performance of different groups" of teachers (AERA et al., 2014, p. 216). If any of the four measures of teacher effectiveness were truly unbiased measures, each of the 48 models should have, in theory, demonstrated no significant associations between any of the predictor variables or covariates and the respective measures of teacher effectiveness.

As discussed above (and also in more depth in Chapter 2), the potential for bias, regardless of the type or “level” of the potentially biasing factor, to be observed in each of the four measures is high. While the technical attributes (e.g., number of performance categories, survey response options) or implicit assumptions (e.g., interval scale that does not function as such) of a measure can certainly shape the distribution of scores (and therefore render them as potentially unreliable, invalid, or biased), there are many non-technical factors in play as well.

As previously noted, such factors include student attributes (e.g., race/ethnicity, poverty, achievement, language ability, dis/ability, age, motivation), teacher attributes (e.g., race/ethnicity, educational background, age, years of experience, socioemotional attributes, subject taught, grade taught), and school attributes (e.g., student body composition, fiscal resource levels, climate or working conditions), among others. Importantly, these factors never occur in isolation. They also frequently directly interact with each another (i.e., as described above regarding potential explanations for bias in teachers’ classroom observations and PPP scores). This multiplicity of possibilities only serves to further complicate the process of accurately assessing whether measures of teacher effectiveness are unbiased (and reliable and valid).

Interpretation 2: True differences in teacher effectiveness. Given the consistency of some of the relationships and directionality between specific predictor variables and measures of teacher effectiveness (e.g., percent of ELL students in a teacher’s school), it is possible that these relationships can be explained by actual teacher quality differences. This suggestion is one that Amrein-Beardsley and Geiger (revise and resubmit) also noted in their findings, and I would be remiss here if I did not note the same.

The quality of teachers (e.g., as defined by current measures of teacher effectiveness, such as VAMs, classroom observations, and SPS; advanced credentials; years of experience; educational attainment or degrees) has tended to vary greatly across different types of schools. For example, lower quality teachers have been overrepresented in schools with higher proportions of FRL and URM students (e.g., Glazerman & Max, 2011; Goldhaber et al., 2018; Mansfield, 2015; Sass et al., 2012; Steele, Pepper, Springer, & Lockwood, 2015). Further, FRL and URM students tend to cluster in high-needs (i.e., Title I) schools (Baker, Farrie, Johnson, Luhm, & Sciarra, 2017; Knight, 2019), which often have fewer resources to serve both teachers and students. Schools with fewer resources are generally less attractive to teachers, as insufficient resources can affect a teacher's perception of a school's overall working conditions (Horng, 2009), among other things (e.g., salary). Poor working conditions are one factor that has been shown to affect the quality of teachers who work in such schools (Johnson, Kraft, & Papay, 2012; Ladd, 2011). Some researchers have also indicated that teachers prefer to work in schools with fewer URM students (e.g., Engel, Jacob, & Curran, 2013) and are more likely to leave schools with higher proportions of URM or FRL students (Clotfelter, Ladd, & Vigdor, 2011; Golderhaber, Gross, & Player, 2011; Hanushek, Kain, & Rivkin, 2004), though it is unknown if that is due to the students themselves, the insufficient resources for the schools, poor working conditions, or the higher likelihood of unfair teacher evaluations based on certain types of students. Further, teachers will fewer years of experience or those who are underqualified—both of whom typically have lower teacher evaluation scores—are more likely to be hired at schools with higher proportions of FRL and URM students (Darling-Hammond, 2004b). Researchers have not yet been able to determine the main driver(s) of this pattern, given the numerous and interrelated factors at play, but the one clear consensus is that both teachers and students are not randomly clustered in schools.

This nonrandom clustering drastically complicates the question of how effective a given teacher is (Jackson, 2014; Paufler & Amrein-Beardsley, 2014; Rothstein, 2009, 2010). While technical concerns potentially affecting each of the four teacher effectiveness measures cannot be ignored, the noted relationships among teacher quality, student demographic factors, and the clustering of both teachers and students in certain schools is worthy of critical examination, especially before trying to determine what might be the more accurate interpretation of this study's results.

A caution about causality. I must make one final note regarding any conclusions that one might draw from this study's results, and that involves circling back to one of the study's main limitations: the inability to ascribe causality. Like in Amrein-Beardsley and Geiger's study, I conducted inferential statistical tests to determine the presence, directionality, and strength of relationships between demographic factors and teacher effectiveness measures. However, due to some of the limitations of this study (i.e., lacking the source code that generated teachers' VAS estimates and additional student-level data, as described previously in Chapter 3), I cannot make any statement about the causality of such significant associations.

The only definitive statement that I can make is that there were significant associations between certain student demographic factors at the school level and certain teacher effectiveness measures, while controlling for specific student demographic factors across teacher's classes and specific teacher demographic/professional background variables. It is not possible to know what the true cause was behind these associations. While either of the above interpretations are possible, the most likely scenario is that both partially explain the significant relationships between the student demographic factors at the school level and teacher effectiveness measures. Even more likely is that the majority of the unexplained variance in each of the four measures can be explained by factors that were either not included in this study, or that are currently unknown. In the next and final chapter, I discuss four possible implications of this study's findings, and conclude the dissertation with thoughts on recommendations for future research and inquiry.

CHAPTER 6

DISCUSSION AND CONCLUSION

In this final chapter, I first provide a brief summary of the study and most relevant results. I then continue with a more in-depth discussion of the implications I inferred from the study's results, from both an applied perspective and a more theoretical perspective. Lastly, I close the dissertation by addressing directions for future research and inquiry.

Study Summary

Since No Child Left Behind (NCLB) and the multi-billion dollar Race to the Top (RTTT) initiative, the focus on holding schools and teachers accountable for their students' achievement has never been more intense or under more scrutiny. Most notably, this focus on accountability led to the proliferation of value-added models (VAMs) being used to evaluate teachers' effectiveness and, in many cases, inform highly consequential personnel decisions. In spite of the ever-growing concerns about the reliability, validity, bias, fairness, and transparency of VAMs, over 80% of states required VAMs to be incorporated in their teacher evaluation systems by 2015 (Doherty & Jacobs, 2015).

Ultimately, the noted statistical and pragmatic concerns about VAMs resulted in over a dozen lawsuits being filed across the country, where teachers contested state or district teacher evaluation policies (see Sawchuck, 2015). One such case was in New Mexico in 2015, where plaintiffs claimed that the state's VAM—which comprised up to 50% of teachers' overall evaluation ratings during the 2013-2014 through 2015-2016 school years—produced unreliable and inaccurate estimates of teachers' levels of effectiveness (*State ex rel. Stewart v. New Mexico Public Education Department*, 2015). These unreliable and inaccurate estimates led to unfair consequences for teachers who were deemed to be ineffective

Related to this lawsuit, I previously examined whether and to what extent the teacher evaluation measures used in New Mexico during the 2013-2014 through 2015-2016 school years showed indications of (un)reliability, (in)validity, and bias (or lack thereof) (Amrein-Beardsley & Geiger, revise and resubmit). While findings from those analyses indicated that the measures used were likely unreliable, invalid, and biased, I never took potential confounding factors into account. As such the purpose of this study was to build upon Amrein-Beardsley and Geiger's prior work, specifically around the notion of bias, with the goal of strengthening their findings and continuing to shed light into the measurement properties previously and currently used to evaluate and hold teachers accountable.

My overarching research questions were 1) What are the relationships between student background characteristics, aggregated to the school level, and the four main teacher evaluation measures that comprised a teacher's overall evaluation score in New Mexico during the 2013-2014, 2014-2015, and 2015-2016 school years? and 2) How do these relationships compare across the four main teacher evaluation measures? To answer these questions, I used multiple linear regression, which allowed me to determine the potential significance of such relationships. I created separate yearly samples of teachers based on the school years for which I had data. Each sample consisted of all public school, non-charter, certified teachers for whom there was course-specific data; who taught at least 15 students across all of their classes; and who had VAS scores, classroom observation scores, PPP scores, and SPS scores.

In the previous chapters, I presented my results (Chapter 4) and findings (Chapter 5) as organized per each measure of teacher effectiveness. I also discussed two possible broad interpretations of the study results in Chapter 5. While I am unable to determine which of the two broad interpretations is more accurate, my discussion of each should have provided the opportunity for sufficient reflection upon the many nuances that can affect teachers' effectiveness scores and ratings. In the next section, I discuss four possible implications drawn from this study's findings.

Implications

Proper Use and Interpretation of Measures

While the teacher evaluation system in New Mexico looks different now than during the 2013-2014 through 2015-2016 school years, as the state no longer uses student growth data (NMPED, 2019b, 2019c), there are still several implications to be drawn from this study that should be of use for both policymakers and practitioners. First, and possibly most importantly, is the oft-stated common implication resulting from research on high-stakes teacher evaluation systems: use teachers' scores solely as intended and interpret results with caution.

To put it succinctly, if a measure of teacher effectiveness shows possible evidence of bias, it begs the question of whether that measure should be used to evaluate teachers. At the very least, such a measure should not be used for summative purposes, at least until independent researchers (e.g., those with no connections to any company that develops, sells, or owns a measure, those with no fiscal or other incentive to carry out the research, etc.) can provide enough evidence that the question of potential bias is laid to rest. For over a decade, this very practice has been cautioned against, specifically with VAMs, as numerous researchers, scholars, and professional organizations warned of attaching such high stakes decisions to potentially misleading or outright faulty outputs (e.g., AERA Council, 2015; Amrein-Beardsley, 2014; ASA, 2014; Braun, 2005; Darling-Hammond, 2015). Yet in many states and districts, those warnings went unheeded and, in the case of New Mexico (among others), teachers filed suit contesting the fairness and accuracy of such measures (*State ex rel. Stewart v. New Mexico Public Education Department*, 2015).

The combination of the passage of the Every Student Succeeds Act (2015); numerous teacher evaluation lawsuits across the country between 2010 and 2015 (see Sawchuk, 2015); teachers' growing discontentment with "extreme" and "out-of-touch" evaluation measures and systems (Burgess, 2017, paras. 21-22); the academic community providing cautionary warnings that increased in both frequency and severity about the ills of VAMs; and the majority of teachers and administrators distrusting VAMs (Harris & Herrington, 2015) likely collectively spurred the decrease in the number of states that currently use VAMs. However, and perhaps surprisingly, VAMs are still used in over a quarter of states (Close et al., 2018; Ross & Walsh, 2019) in spite of the documented measurement controversies, pragmatic concerns, and related lawsuits.

Potentially just as concerning in the context of possible bias, if not more so, has been the increased use and weight of other measures (e.g., classroom observations, SPSs;

Ross & Walsh, 2019) that have the potential to evidence high levels of bias as well. For example, once New Mexico removed the student growth component from its teacher evaluation system, the classroom observation component accounted for the majority of all teachers' overall effectiveness ratings, with teachers' PPP and SPS scores accounting for the rest (NMPED, 2019b, 2019c). While the state's evaluation system is no longer as exclusionary as it was once, since all teachers can be evaluated by all measures (unlike previously, as only 30% or so of teachers could be evaluated by VAMs; Amrein-Beardsley & Geiger, revise and resubmit), this improved fairness by no means signifies that all teachers will be evaluated on a level playing field. Plenty of work remains to ensure that each measure's measurement properties are sound and all teachers have a fair and equal chance to demonstrate their abilities—regardless of how "effectiveness" is defined—irrespective of the students they teach, the schools in which they work, or their own biological characteristics. If nothing else is taken from this study, one clear implication is that all states, districts, administrators, and policymakers should exercise extreme caution when interpreting measures of teacher effectiveness. No measure is devoid of error, regardless of how precise or accurate it might appear to be, or of how well accepted and unquestioned it is or appears to be by policymakers, the academic community, or the media, for example.

Further, as was demonstrated via this study, and others, error in each measure can be exacerbated by confounding factors (e.g., student and teacher demographics; Castellano, Rabe-Hesketh, & Skrondal, 2014). Until the creators and developers of student growth models, classroom observation frameworks, and student surveys—along with any new or additional measures, such as SLOs—can clearly and consistently demonstrate that their respective measures meet the *Standards* (AERA et al., 2014), it would behoove all involved to critically examine and question blanket statements made about a measure’s attributes, quality, or use, especially regarding its performances across multiple factors of student, teacher, classroom, and school contexts. While somewhat clichéd, if something sounds too good to be true, it usually is.

Expansion of Scope of Research

A second implication from this study, which is also related to the first, is the need to expand the scope of teacher effectiveness research as related to the measures used to evaluate teachers. For the past decade, the majority of research about teacher effectiveness during the accountability movement has been about VAMs (Harris & Herrington, 2015. This research was certainly needed, especially given the polemic nature of VAMs (Amrein-Beardsley & Holloway, 2019; Lavery et al., in press) and the potential for misuse and abuse (see, for example, Amrein-Beardsley, 2014). However, the spotlight on VAMs seemed to minimize, if not render completely invisible in some contexts, the measurement and pragmatic concerns related to other measures. Now, with non-VAM measures informing summative decisions more than ever before in the past decade (see Ross & Walsh, 2019), these concerns can no longer be ignored.

While more research is absolutely needed (and likely always will be) about the statistical and methodological properties of all measures used to evaluate teachers, especially if such measures are used wholly or in part for summative purposes, research that also extends beyond these technical properties is desperately warranted. The complexities and interactions among multiple contextual and situational factors (e.g., teacher identities, student identities, school climate) must be taken into consideration (Castellano et al., 2014), or else we run the risk of the legitimacy and the constitutionality of any measure being challenged. If utilizing multiple measures to evaluate teachers continues to be emphasized, this call for the extension of such research becomes even more paramount as combining measures allows for the possibility of a whole host of new measurement and pragmatic concerns.

The current landscape of teacher effectiveness research has generally omitted a full discussion of the myriad of potential influences affecting student achievement and teaching effectiveness measures, respectively. The factors that affect student achievement and teaching effectiveness, respectively, are incredibly complex and multifaceted (Cochran-Smith, 2003). Yet many researchers, as demonstrated by the specifics of their studies (e.g., analytic method, research design), have implicitly assumed that schools and the phenomena within them are not influenced by contextual factors (Skourdoumbis & Gale, 2013). While it is not reasonable for researchers to include literally any and every variable that might affect a given outcome, at minimum they should make concerted efforts to educate consumers of their work about the complexities of such intricate processes (AERA Council, 2011).

Lastly, as related to this implication, researchers who have contributed to the current literature base seem to have completely understated, if not ignored, the seriousness of potential methodological and pragmatic concerns with non-VAM measures (e.g., as outlined above and previously in Chapter 2). While the results of this study indicate that classroom observations, the PPP component, and student surveys might not be as susceptible to school-level bias as VAMs, that by no means implies that those measures are unlikely to be biased by other confounding factors, either at the school level or elsewhere (Castellano et al., 2014). Given the history of teacher accountability policy and preponderance of teachers' effectiveness ratings being informed by VAMs, one can understand why researchers have prioritized examinations into VAMs' technical and pragmatic concerns over those of other non-VAMs measures. However, now that non-VAM measures are becoming the factors weighing most heavily into teachers' effectiveness ratings, it is time for research priorities to be realigned as the documented concerns about each of the other measures can no longer be swept under the proverbial VAM rug.

The "Gold Standard," But at What Cost

A third and final implication from this study, which is more conceptual in nature compared to the first two, is the need to critically question the notion that the distribution of teaching quality should be normally distributed (i.e., a bell curve). As previously described in Chapter 2, most VAMs function by producing an estimate of a teacher's effectiveness (i.e., "added value") relative to other similar teachers' estimates. The normative nature of VAMs results in teacher estimates that are normally distributed. Per the statistical properties of a normal distribution, in order for scores to remain normally distributed, as one teacher's effectiveness rating improves, another teacher's has to decline (see Cattell, 1994; Hicks, 1970)—regardless of how truly effective one or both of the teachers are. Given these statistical properties, at a minimum around two to three percent of teachers will always be calculated has having added “much less value” and “much more value,” respectively, than their peers—regardless of how much “value” each teacher actually adds to his or her students’ learning and achievement.

Ironically, the normative nature of a bell curve works to actually obscure exactly what teacher evaluation systems are trying to accomplish: determine which teachers are effective and which are not. This resultant ambiguity is something that seems to be lost on those who believe that a normal distribution is an accurate representation of teaching quality. Further, and as previously noted, when other measures of teacher effectiveness are forced to “align” with VAM scores (i.e., to produce a more normal distribution of summative teacher effectiveness ratings), the very construct that each measure is intended to assess is invalidated (AERA et al., 2014; see also Amrein-Beardsley & Geiger, 2019b).

Not only is a normal distribution a highly unlikely representation of teaching quality (see Baker et al., 2013; Gould, 1996), but the logic behind the push for a normal distribution of teacher effectiveness ratings is inherently flawed. When Weisberg et al. (2009) argued that their highly negatively skewed distribution of teacher effectiveness ratings was illogical, they used the lack of a highly negatively skewed distribution of student achievement as their rationale. However, they completely failed to recognize the multitude of other known and unknown factors that affect student achievement (e.g., Cochran-Smith, 2003; Kennedy, 2010b).

In New Mexico, as previously noted, many lauded the state’s effort in the push for a normal distribution of teaching quality. It was one of two states that was singled out with praise for having more than 1% of its teachers rated in its lowest effectiveness category (i.e., “ineffective”) (Kraft & Gilmour, 2017), and it was also distinguished as one of six states and districts as a “pioneer” in “reflect[ing] the genuine distribution of teacher talent” (Putnam et al., 2018, p. 3). Putnam et al. labeled New Mexico’s wider (i.e., less skewed) distribution of teacher effectiveness ratings as evidence of the state “making a difference” and its evaluation system as “getting results,” as per the title of their report.

Regardless of the praise that New Mexico received for its so-called “success” in achieving a more normal distribution of teaching quality, their quest for this distribution clearly came at a cost to the state—and to an even larger cost to the individual teachers whose jobs and careers were affected. The individual measures within the NMTEACH system were likely susceptible to bias (as per the results of this study; see also Amrein-Beardsley & Geiger, revise and resubmit), and teachers’ overall effectiveness ratings were relatively inconsistent and unstable (Doan et al., 2019). This calls into question the validity of any inferences drawn from the individual measures and the system as a whole. The dubious nature of such inferences was actually of concern in multiple lawsuits—the *State ex rel. Stewart v. New Mexico Public Education Department* (2015) being one and *Lederman v. King* (2014) being another.

Ultimately, the very system that was so highly touted that it was referred to as a “gold standard” was also the very one that was so damaging and flawed that it was essentially determined to be unconstitutional—at least until the state could prove otherwise (which it was not able to do). It does not seem possible or plausible that any system or state should be labeled as one worth emulating when its measures are flawed and potentially biased to the point of being unconstitutional, and when its overarching goal (i.e., achieving a normal distribution of teaching quality) has no empirical basis and is grounded in faulty logic.

The rationale used to support a normal distribution of teacher quality is eerily parallel to that which was used to argue the true distribution of the people's intelligence (IQs) reflects that of a bell curve. This argument, which was published in the highly influential book, *The Bell Curve* (Herrnstein & Murray, 1994), was extremely controversial due to its evident racist undertones and strongly worded yet generally under-substantiated claims about the nature of intelligence across the human race. At the forefront of the controversy was the presented "fact" that the general achievement gap between white and non-white Americans was, essentially, due to innate biological differences in cognitive ability (i.e., IQ) and related patterns of social behavior.

Herrnstein and Murray argued that those with higher cognitive abilities typically engaged in more socially desirable behaviors while those with lower cognitive were more likely to engage in socially undesirable behaviors, which included things like getting a divorce, bearing children out of wedlock, committing crimes, being unemployed, and dropping out of school. Per Herrnstein and Murray's logic, which some labeled as being steeped in "anachronistic social Darwinism" (Gould, 1995, p. 4), since people of different races were found to have different levels of cognitive abilities, it only made sense that such "undesirable behaviors" were more likely to be committed by people of certain races versus others. Herrnstein and Murray (1994) entrenched themselves in scientific racism (e.g., Graves, 2001) by manipulating or frequently misrepresenting statistics to provide “evidence” as support for their assertions, along with using often contradictory and circumstantial logic (Gould, 1995).

Many critics of the book (e.g., Fraser, 1995; Jacoby & Glauberman, 1995) directly challenged Herrnstein and Murray’s implicit assumptions, (mis)use of statistics, and abject racism. For example, Gould (1995) challenged the assumptions that IQ could be reduced to a single numerical value and that people should be ranked or ordered by such values. He also argued that the data used by Herrnstein and Murray to evidence their claims was “extraordinarily one-dimensional” (p. 7) and the analytical methods they employed were “in violation of all statistical norms” (p. 10). Other critics (e.g., Newby & Newby, 1995) accused Herrnstein and Murray (1994) as implicitly espousing the eugenics movement, while Herbert (1994) classified Herrnstein and Murray’s (1994) writing as “a genteel way of calling somebody” the n-word (para. 6). Regardless of these criticisms, many laypeople, along with some scholars, treated the hand-picked statistics Herrnstein and Murray presented throughout the book as unbiased facts. These so-called facts served, in many cases, to further perpetuate unsubstantiated claims about IQ, cognitive ability, race, and achievement, and many in the media quickly accepted Herrnstein and Murray’s (1994) conclusions without much fanfare or debate (Naureckas, 1995).

The parallels between the rationales to defend normal distributions of both teaching quality and IQ are numerous. Both rationales uphold the assumptions that the latent constructs of “effective teaching” and “intelligence,” respectively, can and should be accurately identified, assessed, quantified, and reduced to single numerical indicators.

Both rationales further assume that different teachers (or groups of teachers) can and should be accurately compared by their teacher effectiveness ratings, and that different people (or groups of people) can and should be accurately compared by their IQs. Further, the idea that normal distributions were the most accurate representations for teaching quality and IQ, and the rationale used to support both claims, were heralded by many policymakers (regarding teaching quality), the media (regarding IQ), and laypeople (regarding both), despite many academic researchers and professional organizations taking strong stands against the accuracy of such distributions, the arguments and logic used to support the cases for such distributions, or both (e.g., AERA Council, 2015; Amrein-Beardsley, 2014; ASA, 2014; Doherty & Jacobs, 2015; Gould, 1995, 1996; Naureckas, 1994; Neisser et al., 1996; Ross & Walsh, 2019; Weisberg et al., 2009).

Nowhere in the literature on teacher effectiveness or in *The Bell Curve* are there even mere acknowledgements—let alone scholarly discussions—of the elitist origins of the normal distribution and related statistical analyses (see, for example, Porter, 1995). There is also minimal to no acknowledgment that standardized tests (which inform both teacher effectiveness ratings and IQs) were created by middle to upper class white men *for* middle to upper class white men. There is also no credence given to the inherent cultural bias in such tests. Further lacking are any understandings of the potential perils of reducing a complex phenomenon like teaching and a multifaceted construct like intelligence to single numerical indicators (Rose, 1991, 1999; Taubman, 2009), or the fact that such reductionistic actions are never apolitical or ideologically neutral (Starr, 1987).

Further, and possibly the most nefarious assumption inherent in each of these respective rationales, is how any measured shortcoming (i.e., a low score on a teacher effectiveness measure or a low overall teacher effectiveness rating; a low IQ) is attributed to the individual who possesses that shortcoming, rather than to the systems and institutions that created and legitimized the mechanisms for such measurements and comparisons in the first place. Psychologists have coined this phenomenon as a "fatal attribution error," which occurs when a person's behavior is overly attributed to his or her personal attributes while simultaneously being under attributed to the contextual or situational effects that might have led to such behavior (Ross, 1977; see also Kennedy, 2010a).

In the case of teacher evaluation, the fatal attribution error is so damaging because it allows the flawed logic that supported punitive accountability policies to flourish. Once teachers were determined to be the biggest in-school influencer on students' achievement (Coleman et al., 1966) and the *A Nation at Risk* report (NCEE, 1983) was released, there was finally somebody to take the blame for the educational crisis of apparently declining test scores that resulted in the fear that the country would lose its reputation as an international superpower.

The fatal attribution error actually works in policymakers' favors, as it allows them to absolve themselves of any responsibility or wrongdoing when policies (e.g., like those to raise student achievement) do not work as intended. Ironically, when policies fail, those in power and control (e.g., policymakers) commit what is essentially a non-fatal attribution error, but onto themselves: they explain away any purposeful or calculated decision that might have played a part in the policy failing as being informed by mere "value-neutral" "technical expertise" (Lingard, 2011; Rose, 1999). This rationalization replaces the admission of responsibility for their roles in and contribution to a failed goal. It also saves them from acknowledging that the failures (or successes) of any policies should be collectively ascribed to all individual actors involved and the systems and contexts in which such policies operate.

Over the past five or so years, the often unquestioned and blindly accepted assumptions and logic used to support a normal distribution of teaching quality have slowly begun to raise suspicions about the extent of their accuracy and rationality. However, these suspicions have been evidenced mostly indirectly, through critiques of the actual instruments, tools, measures, and related practices that are supported by such assumptions and logic (e.g., VAMs, forcing alignment between two measures). Sooner or later, we must begin to question the assumptions and rationale that undergird the systems that decide, far in advance of a teacher ever instructing a student, the specific proportions of teachers who will be calculated as minimally and maximally effective, regardless of how effective those teachers might or might not be.

Conclusion

Ever since researchers documented that teachers were the most influential in-school factor contributing to students' learning and achievement (e.g, Coleman et al., 1966) and federal policies and initiatives mandated that teachers needed to be held accounting for their students' learning and achievement, evaluating teachers' effectiveness has taken on increased importance even though researchers have found that teachers only account for as much as 14% of the variance in students' test scores (see ASA, 2014). However, the overly complex nature of teaching (Cochran-Smith, 2003) and the debate over what artifact best represents student learning and achievement, coupled with the standards and accountability movement, has resulted in more controversy than answered questions. Common longstanding arguments have included debates about whether and to what extent current measures of teacher effectiveness and overall teacher evaluation systems are reliable, valid, unbiased, transparent, and fair.

In the past decade, states' and districts' (over)reliance on VAMs to hold teachers accountable and to inform highly consequential personnel decisions ultimately backfired. Results from this study contribute to the current body of literature on teacher evaluation by providing yet another warning and cautionary tale about the dangers of using potentially biased measures to both evaluate and hold teachers accountable for their students' learning. While states' and districts' reliance on VAMs seem to be finally fading (Close et al., 2019; Ross & Walsh, 2019), the measures that are replacing the weight of VAMs in teacher evaluation systems are not necessarily any more methodologically sound or less susceptible to legal contestation.

While these individual components of teacher evaluation systems are shifting, the accountability movement—including holding teachers accountable for their students' achievement—remains cemented in place. The discourse and rhetoric used to manufacture a so-called crisis (Berliner & Biddle, 1995) about the decline of American education in the early 1980s are still present today, although possibly more subversively so. The main tenets of the logic underpinning the accountability movement and holding teachers accountable have become legitimized to the point where teachers are committing their own fatal attribution errors about themselves (see Ball, 2003; Holloway, 2019). For example, when a teacher receives a low effectiveness score, his or her first instinct tends to be to attribute that score to his or her own lack of ability, rather than to question the very system and logic underpinning that system that makes such a score possible and necessary in the first place (Gabriel, 2017; Holloway & Brass, 2018; Lewis & Holloway, 2018; see also Ball, 2003).

Reframing the paradigm that has discursively positioned teachers as the ones to blame if their students' learning is not clearly made evident takes time, effort, and resources. Such resources include scholars within the academic community who are willing to continue to research the more technical components of the tools and measures used to evaluate teachers, as well as to critically examine and raise questions about the many assumptions inherent in today's evaluation of teachers. Without these lines of inquiry being investigated, it is unlikely that there will be any change in the way teachers are ultimately held accountable, and therefore blamed, for their students' achievement.

Recommendations for Further Study

Although the body of literature on contemporary teacher evaluation systems and measures has increased exponentially since the passage of the RTTT initiative (USDOE, 2009a), researchers need to continue to focus their efforts on these areas given the numerous outstanding concerns about teacher evaluation systems and their measures. In the following subsections, I offer recommendations for future study from a shorter term technical perspective and a longer term conceptual perspective.

Shorter term: Technical perspective. Given the lack of overall consensus on each individual measure's levels of (un)reliability, (in)validity, and bias (or lack thereof), my first recommendation is to repeat this study, albeit with several methodological changes to obtain a more robust result. First I would recommended that additional demographic and background variables that were missing from the NMPED dataset be included. Given the overall low proportions of variance that each model explained in its respective outcome measures, there are invariably additional data points that can—and should—be included in models like the ones analyzed herein.

I would also want to test for possible interaction effects between some of the covariates in the models, as it is quite possible that the effect of one predictor variable (e.g., student race/ethnicity) on the criterion variable (e.g., percent of SPS points a teacher earned) depends on the value of a second predictor variable (e.g., teacher race/ethnicity). I believe this is especially likely for measures that are more subjective in nature, such as classroom observations and SPSs. I would also recommend reconsidering the assumption that the predictor variables and covariates have a linear effect on the criterion variables, as researchers have demonstrated that some variables affecting teacher quality, such as teachers' years of experience (e.g., Boyd, Lankford, Loeb, Rockoff, & Wyckoff, 2008; Harris & Sass, 2011; Ladd & Sorenson, 2017; see also Podolsky et al., 2019), do not have a linear effect.

More important than repeating this study with the above changes is expanding the scope of research (as previously discussed in the Implications section in Chapter 5). As discussed, much of today's current research on teacher effectiveness conceptually undervalues the contributions of contextual factors on student achievement (Kennedy, 2010a). While such omissions are common in policy research that functions within a framework of accountability policy (Skourdoumbis & Gale, 2013), as is currently the case, the reality is that findings and implications from such studies might be completely misrepresented. Further, since current evaluation policies, including high-stakes decisions, have been shaped by such research, these possible omissions are especially important.

I recommend that first, and similar to a prior recommendation, researchers examine the possibilities for interactions among variables that are commonly included in teacher effectiveness analyses. There are many aspects of teachers' and students' identities that, when intersecting, can change the entire dynamic of an interaction or learning opportunity, for example. I recommend that such possible interactions be assessed especially for classroom observations and SPSs, given the numerous actors involved and the potential for a variety of different unconscious biases to affect these evaluations. Depending on the measure of teacher effectiveness being assessed, examples of such interactions might be the intersection of common demographic factors, like student gender x teacher gender or observer race/ethnicity x teacher race/ethnicity. These interactions should also be examined across multiple types of classrooms (e.g., mathematics, reading, special education), grade/school levels (e.g., elementary, middle, high school), and schools (e.g., high needs/Title I schools, more affluent schools), as is already common.

While the above recommendation regarding expanding the scope of research is important, the main expansion that needs to occur is that into the often overlooked and underappreciated situational or contextual aspects of schooling and teachers' working conditions that likely affect, on some level, teachers' effectiveness. These aspects are neatly summarized and organized into four broad categories by Kennedy (2010a): a) the parameters of the work itself, b) the students in a teacher's classroom and school, c) institutional practices that interrupt classroom life, and d) policy reforms. For the purposes of brevity, rather than discuss in detail each of these four categories here, I urge those interested to see Appendix D, review Kennedy's (2010a) discussion of these important factors, and subsequently include such aspects in future empirical or conceptual work.

Longer term: Conceptual perspective. While empirical analyses like those mentioned above are without a doubt beneficial to teachers, students, and the discipline of education writ large, I also offer recommendations regarding several of the assumptions that most in the education community in the U.S. seem to have taken for granted, blindly accepted, or not yet realized. Until such assumptions begin to be made transparent and then critically questioned, it is likely that teachers will continue to bear the brunt of the responsibility for their students' achievement while being assessed by measures that are methodologically unsound, irrespective of how fair or unfair that may be.

I first urge researchers and interested readers to critically examine and discuss not just *how* to best define the construct of "teacher effectiveness," but *why* that definition is more valid than others. I recommend the same exercise for how to best define the notion of "student achievement," as well as what the actual distribution of teacher quality is and what the ideal distribution of teacher quality should be.

I also recommend that the trend of numericization (Rose, 1991, 1999)—transforming a social practice, like teaching, into numbers—be critically questioned. Specifically, I urge that researchers and laypeople alike specifically question the underlying assumptions that numerical indicators (e.g., a teacher's overall effectiveness rating) are objective, apolitical, value-neutral, and accurate, and should be prioritized over other (i.e., non-numerical) forms of data (Dixon-Roman, 2016; Gould, 1996; Lingard, 2011; Poovey, 1998; Porter, 1995; Rose, 1999). Related to the fatal attribution error, I urge all, but especially teachers, to question why failing to provide evidence of quantifiable student achievement is solely attributed back to individual teacher, but especially in the form of a personal value judgment (see Harvey, 2007; Holloway & Brass, 2018; Lewis & Holloway, 2018), as opposed to being attributed to *all* factors that affect student achievement in the first place.

While I am certain that many additional questions along similar lines exist, my goal in making the above recommendations for further inquiry is to, at the very least, shed light on a small fraction of the many implicit and often taken for granted assumptions that are inherent in teacher evaluation systems and policies. When such assumptions remain unchecked, they serve to further support and perpetuate current policies that result in teacher evaluation practices that are both unfair and inaccurate—as was evidenced in New Mexico, among other states. These questions and the resulting critical examinations that are (hopefully) likely to occur are the only way to make many of the dangerous assumptions used to undergird the logic of the teacher accountability movement transparent, which is a necessary precursor to widespread change. Until the current hidden assumptions become at transparent, the likelihood that teachers will continue to be assessed via measures that are unreliable, invalid, or biased seems unfortunately high.

REFERENCES

A-B-C-D-F Schools Rating System. S.B. 427, 50th Leg. (1st session). (2011).

Adler, M. (2013). Findings vs. interpretation in "The long-term impacts of teachers" by Chetty et al. *Education Policy Analysis Archives, 21*(10). doi: 10.145/07/epaa.v21n10.2013

Adler, M. (2014). Review of *Measuring the impacts of teachers*. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/files/ttr-chetty-teachimpacts_0.pdf

Albrecht, S. F., & Joles, C. (2003). Accountability and access to opportunity: Mutually exclusive tenets under a high-stakes testing mandate. *Preventing School Failure, 47*(2), 86-91. doi: 10.1080/10459880309604435

Aldeman, C., Hyslop, A., Marchitello, M., Schiess, J. O., & Pennington, K. (2017). *An independent review of ESSA state plans*. Sudbury, MA: Bellwether Education Partners.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*(2), 153-166. doi: 10.1023/A:1008168421283

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*(6045), 1034-1037. doi: 10.1126/science.1207998

Alm, B. J. (2017). *Giving voice to teachers: How twenty 4th and 5th grade teachers describe and understand experiences receiving value-added scores in New York* (Doctoral dissertation). Retrived from ProQuest. (10621328)

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Educational Research Association (AERA) Council. (2011). *AERA code of ethics*. Washington, DC: Author.

American Educational Research Association (AERA) Council. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. Educational Researcher, *44*(8), 448-452. doi:10.3102/0013189X15618385

American Federation of Teachers. (2015). *Teachers and state legislators join AFT New Mexico in lawsuit challenging constitutionality of punitive, error-ridden teacher evaluation system*. Albuquerque, NM: Author.

American Recovery and Reinvestment Act (2009), P.L. 111-5, 123 Stat 115.

American Statistical Association (ASA). (2014). *ASA statement on using value-added models for educational assessment*. Alexandria, VA.

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability*. New York, NY: Routledge.

Amrein-Beardsley, A. (2016a). Category archives: Lawsuit [Web log]. Retrieved from http://vamboozled.com/category/lawsuit/

Amrein-Beardsley, A. (2016b, February 18). Chetty et al. v. Rothstein on VAM-based bias, again. Retrieved from http://vamboozled.com/chetty-et-al-v-rothstein-on-vam-based-bias-again/

Amrein-Beardsley, A. (2017, August 14). The “widget effect” report revisited [Web log post]. *Vamboozled!* Retrieved from http://vamboozled.com/the-widget-effect-report-revisited/

Amrein-Beardsley, A. (2018a). *Affidavit of Audrey Amrein-Beardsley* in *State ex rel. Stewart v. New Mexico Public Education Department* (2015).

Amrein-Beardsley, A. (2018b, September 18). Learning from what doesn’t work in teacher evaluation [Web log post]. *Vamboozled!* Retrieved from http://vamboozled.com/learning-from-what-doesnt-work-in-teacher-evaluation/

Amrein-Beardsley, A., & Close, K. (2019). Teacher-level value-added models on trial: Empirical and pragmatic issues of concern across five court cases. *Education Policy*. doi: 10.1177/0895904819843593

Amrein-Beardsley, A., & Geiger, T. (revise and resubmit). Using test scores to evaluate and hold school teachers accountable in the U.S. *Educational Assessment, Evaluation and Accountability.*

Amrein-Beardsley, A., & Geiger, T. (under re-review). Methodological concerns about the Education Value-Added Assessment System (EVAAS): Validity, reliability, and bias. *SAGE Open.*

Amrein-Beardsley, A., & Geiger, T. (2019a, December, 13). New Mexico lawsuit: Final update. *Vamboozled!* Retrieved from http://vamboozled.com/new-mexico-lawsuit-final-update/

Amrein-Beardsley, A., & Geiger, T. (2019b). Potential sources of invalidity when using teacher value-added and principal observational estimates: Artificial inflation, deflation, and conflation. *Educational Assessment, Evaluation and Accountability, 31*(4), 465-493. doi: 10.1007/s11092-019-09311-w

Amrein-Beardsley, A., & Holloway, J. (2019). Value-added models for teacher evaluation and accountability: Commonsense assumptions. *Educational Policy, 33*(3), 516-542. doi: 10.1177/0895904817719519

Amrein-Beardsley, A., & Holloway-Libell, J., Cirell, A. M., Hays, A., & Chapman, K. (2015). “Rational” observational systems of educational accountability and reform. *Practical Assessment, Research & Evaluation, 20*(17). Retrieved from https://scholarworks.umass.edu/pare/vol20/iss1/17

Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist’s companion.* Princeton, NJ: Princeton University Press.

Astor, E. (2018). *Teachers “talk back”: Exploring the dynamics between practice and value-added evaluation policy*. Retrieved from ProQuest. (10974602)

Babad, E. (1993). Teachers’ differential behavior. *Educational Psychology Review, 5*(4), 347-376. doi: 10.1007/bf01320223

Babad, E., Bernieri, F., & Rosenthal, R. (1991). Students as judges of teachers’ verbal and non-verbal behavior. *American Educational Research Journal, 28*(1), 211-234. doi: 10.2307/1162885

Bailey, J., Bocala, C., Shakman, K., & Zweig, J. (2016). *Teacher demographics and evaluation: A descriptive study in a large urban district* (REL 2017-189). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Baker, B. D. (2013, June 10). Revisiting the Chetty, Rockoff, & Friedman molehill [Web log post]. Retrieved from http://nepc.colorado.edu/blog/revisiting-chetty-rockoff-friedman-molehill

Baker, B. D., Farrie, D., Johnson, M., Luhm, T., & Sciarra, D. G. (2017). *Is school funding fair? A national report card* (6th ed.). Newark, NJ: Education Law Center.

Baker, B. D., Oluwole, J. O., & Green, P. C. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era. *Education Policy Analysis Archives, 21*(5). doi: 10.14507/epaa.v21n5.2013

Baker, E. L. (2003). Multiple measures: Toward tiered systems. *Educational Measurement: Issues and Practice, 22*(2), 13-17. doi: 10.1111/j.1745-3992.2003.tb00123.x

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers* (Briefing Paper #278). Washington, DC: Economic Policy Institute.

Balch, R. (2016). Using student surveys at the elementary and secondary levels. In J. A. Grissom & P. Youngs (Eds.), *Improving teacher evaluation systems: Making the most of multiple measures* (pp. 77-88). New York, NY: Teachers College Press.

Ball, S. J. (2003). The teacher’s soul and the terrors of performativity. *Journal of Education Policy, 18*(2), 215-228. doi: 10.1080/0268093022000043065

Ballou, D. (2012). Review of *Measuring the impacts of teachers*. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/files/TTR-ValAdd-NBER.pdf

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65. doi: 10.3102/10769986029001037

Ballou, D., & Springer, M. (2015). Using student test scores to measure teacher performance: Some problems in the design and implementation of evaluation systems. *Educational Researcher, 44*(2), 77-86. doi: 10.3102/0013189X14474904

Basow, S. A., & Montgomery, S. (2005). Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education, 18*(2), 91-106. doi: 10.3200/JOEB.84.1.40-46

Basow, S. A., Phelan, J. E., & Capotosto, L. (2006). Gender patterns in college students’ choices of their best and worst professors. *Psychology of Women Quarterly, 30*(1), 25-35. doi: 10.1111/j.1471-6402.2006.00259.x

Bates, L. A., & Glick, J. E. (2013). Does it matter if teachers and schools match the student? Racial and ethnic disparities in problem behaviors. *Social Science Research, 42*(5), 1180-1190. doi: j.ssresearch.2013.04.005

Beaton, A. E., Rogers, A. M., Gonzalez, E., Hanly, M. B., Solstad, A. Rust, K. F. … Jia, Y. (2011). *The NAEP primer* (NCES 2011-463). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Becker, G. (1964). *Human capital: A theoretical and empirical analysis with special reference to education*. New York, NY: Columbia University Press.

Behuniak, P. (2003). Education assessment in an era of accountability. In J. E. Wall & G. R. Walz (Eds.), *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 335-347). Greensboro, NC: CAPS Press.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2-3), 62-87. doi: 10.1080/10627197.2012.715014

Bell, C. A., Jones, N. D., Qi, Y., & Lewis, J. M. (2018). Strategies for assessing classroom teaching: Examining administrator thinking as validity evidence. *Educational Assessment, 23*(4), 229-249. doi: 10.1080/10627197.2018.1513788

Bem, S. L. (1993). *The lenses of gender: Transforming the debate on sexual inequality*. New Haven, CT: Yale University Press.

Bem, S. L. (1995). Dismantling gender polarization and compulsory heterosexuality: Should we turn the volume down or up? *The Journal of Sex Research, 32*(4), 329-334. doi: 10.1080/00224499509551806

Benham, B. J. (1981). CBTE: Another educational edifice built on quicksand. *The Teacher Educator*, *17*(1), 26-29. doi: 10.1080/08878738109554780

Bennett, W. J. (2012, January 6). The lasting impact of good teachers. *CNN*. Retrieved from http://www.cnn.com/2012/01/11/opinion/bennett-good-teachers/index.html

Berliner, D. C. (2006). Our impoverished view of education reform. *Teachers College Record, 108*(6).

Berliner, D. C. (2009). *Poverty and potential: Out-of-school factors and school success*. Tempe, AZ and Boulder, CO: Education Policy Research Unit & Education and the Public Interest Center.

Berliner, D. C. (2013). Effects of inequality and poverty vs. teachers and schooling on America’s youth. *Teachers College Record, 115*(12).

Berliner, D. C. (2014). Exogenous variables and value-added assessments: A fatal flaw. *Teachers College Record, 116*(1).

Berliner, D., & Biddle, B. (1995). *The manufactured crisis: Myths, fraud, and the public attack on America's public schools*. New York, NY: Perseus Books.

Berry, W., & Sanders, W. (2000). *Understanding multivariate research: A primer for beginning social scientists*. New York, NY: Routledge.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Education Measurement: Issues and Practice, 28*(4), 42-51. doi: 10.1111/j.1745-3992.2009.00161.x

Bill & Melinda Gates Foundation. (n.d.). Frequently asked questions. Retrieved from http://k12education.gatesfoundation.org/teacher-supports/teacher-development/measuring-effective-teaching/why-met-additional-resources/frequently-asked-questions/

Bill & Melinda Gates Foundation. (2012). *Asking students about teaching: Student perception surveys and their implementation*. Seattle, WA: Author.

Birman, B. (2013). *A Nation at Risk*'s policy legacy. Washington, DC: American Institutes for Research. Retrieved from http://www.air.org/resource/three-decades-education-reform-are-we-still-nation-risk#Birman2

Birman, B., Le Floch, K. C., Klekota, A., Ludwig, M., Taylor, J., Walters, K., … O'Day, J. (2007). *Evaluating teacher quality under No Child Left Behind*. Santa Monica, CA: Author.

Bitler, M., Corcoran, S., Domina, T., & Penner, E. (2018). *Teacher effects on student achievement and height: A cautionary tale* (Working Paper No. 26480). Cambridge, MA: National Bureau of Economic Research.

Blazar, D., & Kraft, M. A. (2015). *Teacher and teaching effects on students' academic behaviors and mindsets* (Working Paper No. 41). Washington, DC: Mathematica Policy Research.

Blazar, D., Litke, E., & Barmore, J. (2016). What does it mean to be ranked a ''high'' or ''low'' value-added teacher? Observing differences in instructional quality across districts. *American Educational Research Journal, 53*(2), 324–359. doi: 10.3102/0002831216630407

Borman, G. D., & Kimball, S. M. (2005). Teacher quality and education equality: Do teachers with higher standards-based evaluation ratings close student achievement gaps? *Elementary School Journal, 106*(1), 3-20. doi: 10.1086/496904

Boser, U. (2012). *Race to the Top: What have we learned from the states so far?* Washington, DC: Center for American Progress.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008). *Who leaves? Teacher attrition and student achievement* (Working Paper No. 14022). Cambridge, MA: National Bureau of Economic Research.

Boyd, D., Lankford, H., Loeb, S., Rockoff, J. E., & Wyckoff, J. (2008). The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. *Journal of Policy Analysis and Management, 27*(3), 793-818. doi: 10.1002/pam.20377

Bracey, G. (2007, May 1). Value subtracted: A "debate" with William Sanders [Web log post]. *Huffington Post*. Retrieved from http://www.huffingtonpost.com/gerald-bracey/value-subtracted-a-debate_b_47404.html

Braun, H. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service.

Brennan, R. L. (2013). Commentary on "Validating interpretations and uses of test scores." *Journal of Educational Measurement, 50*(1), 74-83. doi: 10.1111/jedm.12001

Briggs, D. & Domingue, B. (2011). Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District Teachers by the *Los Angeles Times*. Boulder, CO: National Education Policy Center (NEPC).

Brookhart, S. M. (2009). The many meanings of "multiple measures." *Educational Leadership, 67*(3), 6-12.

Brophy, J. (1973). Stability of teacher effectiveness. *American Educational Research Journal, 10*(3), 245–252. doi: 10.2307/1161888

Brophy, J. E., Coulter, C. L., Crawford, W. J., Evertson, C. M., & King, C. E. (1975). Classroom observation scales: Stability across time and context and relationships with student learning gains. *Journal of Educational Psychology, 67*(6), 873-881. doi: 10.1037//0022-0663.67.6873

Brophy, J. E., & Good, T. L. (1986). Teacher behavior and student achievement. In M. E. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.) (pp. 328-375). New York, NY: Macmillan.

Broudy, H. (1972). *A critique of performance based teacher education*. Washington, DC: American Association of Colleges for Teacher Education.

Burgess, K. (2016, September 16). Number of effective teachers keeps dropping. *The Albuquerque Journal*. Retrieved from https://www.abqjournal.com/846826/nm-teacher-evals-number-of-effective-teachers-keeps-dropping.html

Burgess, K. (2017, July 6). Expert: NM teacher evals are toughest in the nation. *The Albuquerque Journal*. Retrieved from https://www.abqjournal.com/1029370/expert-nm-teacher-evals-toughest-in-us.html

Butrymowicz, S. (2012, May 3). Student surveys may help rate teachers. *The Washington Post*. Retrieved from https://www.washingtonpost.com/local/education/student-surveys-may-help-rate-teachers/2012/05/11/gIQAN78uMU_story.html?utm_term=.05de45df7cf8

Campbell, D. T. (1976). *Assessing the impact of planned social change*. Hanover, NH: The Public Affairs Center, Dartmouth College. Retrieved from http://portals.wi.wur.nl/files/docs/ppme/Assessing_impact_of_planned_social_change.pdf

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal, 55*(6), 1233-1267. doi: 10.3102/0002831218776216

Carter, P. (2003). “Black” cultural capital, status positioning, and schooling conflicts for low-income African American youth. *Social Problems, 50*(1), 136-155. doi: 10.1525/sp.2003.50.1.136

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*(5), 757-783. doi: 10.1177/0013164413486987

Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons. *Journal of Educational and Behavioral Statistics, 39*(5), 333-367. doi: 10.3102/1076998614547576

Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review, 51*(5), 292-303. doi: 10.1037/h0057299

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44*(5), 495-518.

Chang, D. F., & Sue, S. (2003). The effects of race and problem type on teachers' assessments of student behavior. *Journal of Consulting and Clinical Psychology, 71*(2), 235-242. doi: 10.1037/0022-006x.71.2.235

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh public schools* (REL 2014-024). Calverton, MD: Regional Educational Laboratory Mid-Atlantic.

Charter Schools Act of 2006. New Mex. Stat. § 22-8B (2007).

Chester, M. D. (2003). Multiple measures and high-stakes decisions: A framework for combining measures. *Educational Measurement: Issues and Practice, 22*(2), 32–41. doi:10.1111/j.1745-3992.2003.tb00126.x

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Discussion of the American Statistical Association's statement (2014) on using value-added models for educational assessment. *Statistics and Public Policy, 1*(1), 111-113. doi: 10.1080/2330443X.2014.955227

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). *Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. American Economic Review, 104*(9), 2593-2632. doi: 10.3386/w19424

Close, K., Amrein-Beardsley, A., & Collins, C. (2018). *State-level assessments and teacher evaluation systems after the passage of the Every Student Succeeds Act: Some steps in the right direction*. Boulder, CO: National Education Policy Center.

Close, K., Amrein-Beardsley, A., & Collins, C. (2019). Mapping America's teacher evaluation plans under ESSA. *Phi Delta Kappan, 101*(2), 22-26. doi: 10.1177/0031721719879150

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2011). Teacher mobility, school segregation, and pay-based policies to level the playing field. *Education Finance and Policy, 6*(3), 399-438. doi: 10.1162/edfp_a_00040

Cochran-Smith, M. (2003). The unforgiving complexity of teaching: Avoiding simplicity in the age of accountability. *Journal of Teacher Education, 54*(1), 3-5. doi: 10.1177/0022487102238653

Cochran-Smith, M. (2005). No Child Left Behind: 3 years and counting. *Journal of Teacher Education, 56*(2), 99-103. doi: 10.1177/0022487104274435

Cochran-Smith, M. (2008). The new teacher education in the Unites States: Directions forward. *Teachers and Teaching: Theory and Practice, 14*(4), 271-282. doi: 10.1080/13540600802037678

Cody, C. A., McFarland, J., Moore, J. E., & Preston, J. (2010). *The evolution of growth models.* Raleigh, NC: Public Schools of North Carolina.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*(6), 378-387. doi: 10.3102/0013189X16659442

Cole, R. W. (1979). Minimum competency tests for teachers: Confusion compounded. *Phi Delta Kappan, 61*(4), 233.

Coleman, J. S., Kelly, D. L., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity* (Report No. OE-38001). Washington, DC: U.S. Department of Health, Education, and Welfare.

Collins, C. (2014). Houston, we have a problem: Teachers find no value in the SAS Education Value-Added Assessment System (EVAAS®). *Education Policy Analysis Archives, 22*(98). doi: 10.14507/epaa.v22.1594

Collins, C., & Amrein-Beardsley, A. (2014). Putting student growth and value-added models on the map: A national overview. *Teachers College Record, 116*(1).

Connally, K., & Tooley, M. (2016). *Beyond ratings: Re-envisioning state teacher evaluation systems as tools for professional growth*. Washington, DC: New America.

Corcoran, S. P. (2010). *Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teaching effectiveness in policy and practice*. Providence, RI: Annenberg Institute for School Reform at Brown University.

Corcoran, S. P., & Goldhaber, D. (2013). Value added and its uses: Where you stand depends on where you sit. *Education Finance and Policy, 8*(3), 418-434. doi: 10.1162/edfp_a_00104

Cornell, S. (1996). The variable ties that bind: Content and circumstance in ethnic processes. *Ethnic and Racial Studies, 19*(1996), 265-289. doi: 10.1080/01419870.1996.9993910

Creemers, B., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice, and theory in contemporary schools*. New York, NY: Routledge.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302. doi: 10.1037/h0040957

Curby, T. W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Mooney, L., Downer, J., … Pianta, R. C. (2011). Within-day variability in the quality of classroom interactions during third and fifth grade. *The Elementary School Journal, 112*(1), 16-37. doi: 10.1086/660682

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C. (2013). General questions about teacher evaluation. Retrieved from http://www.danielsongroup.org/general-questions-about-teacher-evaluation/

Danielson, C. (2016, April 18). Charlotte Danielson on rethinking teacher evaluation. *Education Week, 35*(28). Retrieved from http://www.edweek.org/ew/articles/2016/04/20/charlotte-danielson-on-rethinking-teacher-evaluation.html

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1). doi: 10.14507/epaa.v8n1.2000

Darling-Hammond, L. (2004a). From "separate but equal" to "no child left behind": The collision of new standards and old inequalities. In D. Meier & G. Wood (Eds.), *Many children left behind: How the No Child Left Behind Act is damaging our children and our schools* (pp. 3-32). Boston, MA: Beacon.

Darling-Hammond, L. (2004b). Inequality and the right to learn: Access to qualified teachers in California's public schools. *Teachers College Record, 106*(10).

Darling-Hammond, L. (2012). *Creating a comprehensive system for evaluating and supporting effective teaching*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. New York, NY: Teachers College Press.

Darling-Hammond, L. (2015). Can value-added add value to teacher evaluation? [Commentary]. *Educational Researcher, 44*(2), 132-137. doi: 10.3102/0013189X15575346

Darling-Hammond, L., & Berry, B. (1988). *The evolution of teacher policy*. Santa Monica, CA: RAND Corporation.

Darling-Hammond, L., & Schlan, E. (1992). Policy and supervision. In C. D. Glickman (Ed.)., *Supervision in transition* (pp. 7-29). Alexandria, VA: Association for Supervision and Curriculum.

Darling-Hammond, L., & Wise, A. E. (1985). Beyond standardization: State standards and school improvement. *The Elementary School Journal, 85*(3), 315-336. doi: 10.1086/461408

Darling-Hammond, L., & Youngs, P. (2002). Defining "highly qualified teacher": What does "scientifically based research" actually tell us? *Educational Researcher, 31*(9), 13-25. doi: 10.3102/0013189x031009013

De Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behavior. *Learning Environments Research, 4*(1), 51-85. doi: 10.1023/A:1011402608575

den Brok, P., Brekelmans, M., & Wubbels, T. (2006). Multilevel issues in research using students' perceptions of learning environments: The case of the Questionnaire on Teacher Interaction. *Learning Environments Research, 9*(3), 199-213. doi: 10.1007/s10984-006-9013-9

Derringer, P. (2010). RTT in Tennessee: Assessment done right. *Technology and Learning, 31*(1), 40.

Desimone, L. M., Smith, T. M., & Frisvold, D. E. (2010). Survey measures of classroom instruction: Comparing student and teacher reports. *Educational Policy, 24*(2), 267-329. doi: 10.1177/0895904808330173

Desimone, L. M., Stornaiuolo, A., Flores, N., Pak, K., Edgerton, A., Nichols, T. P., … Porter, A. (2019). Success and challenges of the "new" College- and Career-Ready Standards: Seven implementation trends. *Education Researcher, 48*(3), 167-178. doi: 10.3102/0013189X19837239

Dixon-Roman, E. J. (2016). Diffractive possibilities: Cultural studies and quantification. *Transforming Anthropology, 24*(2), 157-167. doi: 10.1111/traa.12074

Doan, S., Schweig, J. D., & Mihaly, K. (2019). The consistency of composite ratings of teacher effectiveness: Evidence from New Mexico. *American Educational Research Journal, 56*(6), 2116-2146. doi: 10.3102/0002831219841369

Doherty, K. M., & Jacobs, S. (2013). *State of the states 2013. Connect the dots: Using evaluations of teacher effectiveness to inform policy and practice*. Washington, DC: National Council on Teacher Quality.

Doherty, K. M., & Jacobs, S. (2015). *State of the states: Evaluating teaching, leading and learning*. Washington, DC: National Council on Teacher Quality.

Doran, H. C., & Izumi, L. T. (2004). *Putting education to the test: A value-added model for California*. San Francisco, CA: Pacific Research Institute.

Dorans, N. J., & Cook, L. L. (2016). *Fairness in educational assessment and measurement*. New York, NY: Routledge.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82*(1), 62-68. doi: 10.1037//0022-3514.82.1.62

Downer, J. T., Stuhlman, M., Schweig, J., Martinez, J. F., & Ruzek, E. (2015). Measuring effective teacher-student interactions from a student perspective: A multi-level analysis. *Journal of Early Adolescence, 35*(5-6), 722-758. doi: 10.1177/0272431614564059

Driscoll, A., Peterson, K. D., Crow, N., & Larson, B. (1985). Student reports for primary teacher evaluation. *Educational Research Quarterly, 9*(3), 43-50.

Duncan, A. (2012). Change is hard: Remarks of U.S. Secretary of Education Arne Duncan at Baltimore County teachers convening. Washington, DC: U.S. Department of Education.

Eckert, J. M., & Dabrowski, J. (2010). Should value-added measures be used for performance pay? *Phi Delta Kappan, 91*(8), 88-92. doi: 10.1177/003172171009100821

Education Commission of the States. (1983). *Action for excellence: A comprehensive plan to improve our nation's schools*. Denver, CO: Author.

Education Commission of the States. (2004). *ECS report to the nation: State implementation of the No Child Left Behind Act, respecting diversity among states*. Denver, CO: Author.

Education Improvement Act. Tenn. Ch. 535. S.B. 1231/H.B. 752. (1992).

Educational Research Service. (1988). *Teacher evaluation: Practices and procedures*. Arlington, VA: Author.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2016). Selecting growth measures for use in school evaluation systems: Should proportionality matter? *Educational Policy, 30*(3), 465-500. doi: 10.1177/0895904814557593

Ehrenberg, R. G., & Brewer, D. J. (1994). Do school and teacher characteristics matter? Evidence from high school and beyond? *Economics of Education Review, 13*(1), 1-17. doi: 10.1016/0272-7757(94)90019-1

Elam, S. (1971). *Performance based teacher education: What is the state of the art?* Washington, DC: American Association of Colleges for Teacher Education.

Elementary and Secondary Education Act of 1965, Pub. L. No. 89-10, 79 Stat. 27 (1965).

Ellett, C. D., & Teddlie, C. (2003). Teacher evaluation, teacher effectiveness, and school effectiveness: Perspectives from the USA. *Journal of Personnel Evaluation in Education, 17*(1), 101-128.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: The Guilford Press.

Engel, M., Jacob, B. A., & Curran, F. C. (2014). New evidence on teacher labor supply. *American Educational Research Journal, 51*(1), 36-72. doi: 10.3102/0002831213503031

Every Student Succeeds Act of 2015, Pub. L. No. 114-95, Stat. 1177 (2015).

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29*, 1-29. doi: 10.1016/j.learninstruc.2013.07.001

Ferguson, R. F. (2008). *The Tripod project framework.* Cambridge, MA: Harvard University.

Ferguson, R. F. (2010). *Student perceptions of teaching effectiveness*. Boston, MA: National Center for Teacher Effectiveness and the Achievement Gap Initiative, Harvard University.

Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan, 94*(3), 24-28.

Ferguson, R. F., & Danielson, C. (2014). How Framework for Teaching and Tripod 7Cs evidence distinguish key components of effective teaching. In. T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project* (pp. 98-143). San Francisco, CA: Jossey-Bass.

Firestone, W. A. (2014). Teacher evaluation policy and conflicting theories of motivation. *Educational Researcher, 43*(2), 100-107. doi: 10.3102/0013189X14521864

Follman, J. (1992). Secondary school students' ratings of teacher effectiveness. *The High School Journal, 75*(3), 168-178.

Follman, J. (1995). Elementary public school pupil rating of teacher effectiveness. *Child Study Journal, 25*(1), 57-78.

Frankenberg, E. (2013). The role of residential segregation in contemporary school segregation. *Education and Urban Society, 45*(5), 548-570. doi: 10.1177/003124513486288

Fraser, S. (Ed.) (1995). *The bell curve wars: Race, intelligence, and the future of America*. New York, NY: Basic Books.

Frosch, D. (2013, December 17). New Mexico teachers resist a state official's plan for evaluating them. *New York Times*. Retrieved from http://www.nytimes.com/2013/12/18/us/a-push-for-teacher-accountability-meets-resistance-in-new-mexico.html?_r=0

Fuller, E. (2014). *An examination of Pennsylvania school performance profile scores* (Policy Brief 2014-1). University Park, PA: Center for Evaluation and Education Policy Analysis.

Furlong, J., Cochran-Smith, M., & Brennan, M. (Eds.). (2009). *Policy and politics in teacher education: International perspectives*. New York, NY: Routledge.

Gabriel, R. E. (2017). Constructing teacher effectiveness in policymaking conversations. In J. N. Lester, C. R. Lochmiller, & R. E. Gabriel (Eds.), *Discursive perspectives on education policy and implementation* (pp. 219-239). New York, NY: Palgrave Macmillan.

Gabriel, R. E. (2018). Reframing observation: Create a culture of learning with teacher evaluation. *The Learning Professional, 39*(4), 46-49.

Gabriel, R. E., & Lester, J. N. (2013a). Sentinels guarding the grail: Value-added assessment and the quest for education reform. *Education Policy Analysis Archives*, *21*(9). doi: 10.14507/epaa.v21n9.2013

Gabriel, R. E., & Lester, J. N. (2013b). The romance quest of education reform: A discourse analysis of the LA Times' reports on value-added measurement teacher effectiveness. *Teachers College Record, 115*(2).

Gabriel, R. E., & Woulfin, S. (2017). *Making teacher evaluation work: A guide for literacy teachers and leaders*. Portsmouth, NH: Heineman.

Gallagher, H. A. (2004). Vaughn Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education*, *79*(4), 79-107. doi: 10.1207/s15327930pje7904_5

Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis, 37*(2), 224-242. doi: 10.3102/0162373714537551

Geiger, T. J., & Amrein-Beardsley, A. (2019). Student perception surveys for K-12 teacher evaluation in the United States: A survey of surveys. *Cogent Education, 6*(1). doi: 10.1080/2331186X.2019.1602943

Gershenson, S., Holt, S. B., & Papageorge, N. W. (2015). Who believes in me? The effect of student-teacher demographic match on teacher expectations. *Economics of Education Review, 52,* 209-224. doi: 10.1016/j.econedurev.2016.03.002

Gill, B., Bruch, J., Booker, K. (2013). *Using alternative student growth measures for evaluating teacher performance: What the literature says*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic, Institute of Education Sciences, U.S. Department of Education.

Gill, B., Shoji, M., Coen, T., & Place, K. (2016). *The content, predictive power, and potential bias in five widely used teacher observation instruments* (REL 2017-191). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Glass, G. V. (1974). A review of three methods of determining teacher effectiveness. In H. J. Walberg (Ed.), *Evaluating educational performance* (pp. 11-32). Berkeley, CA: McCutchan.

Glazerman, S. M., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D. O., & Whitehurst, G. J. (2011). *Passing muster: Evaluating teacher evaluation systems.* Washington, DC: Brown Center on Education Policy at Brookings.

Glazerman, S. M., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added*. Washington, DC: Brown Center on Education Policy at Brookings.

Glazerman, S. M., & Max, J. (2011). *Do low-income students have equal access to the highest-performing teachers?* (NCEE 2011-4016). Washington, DC: National Center for Education Evaluation and Regional Assistance.

Glazerman, S. M., & Potamites, L. (2011). *False performance gains: A critique of successive cohort indicators.* Washington, DC: Mathematica Policy Research.

Goals 2000: Educate America Act of 1994, Pub. L. No. 103-227 (1994).

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Goe, L., & Croft, A. (2009). *Methods of evaluating teacher effectiveness*. Washington, DC: National Comprehensive Center for Teacher Quality.

Goldhaber, D., & Chaplin, D. D. (2015). Assessing the "Rothstein Falsification Test": Does it really show teacher value-added models are biased? *Journal of Research on Educational Effectiveness, 8*(1), 8-34. doi:10.1080/19345747.2014.978059

Goldhaber, D., Gabele, B., & Walch, J. (2012). *Does the model matter? Exploring the relationship between different student achievement-based teacher assessments*. Seattle, WA: Center for Education Data and Research.

Goldhaber, D., Gross, B., & Player, D. (2011). Teacher career paths, teacher quality, and persistence in the classroom: Are public schools keeping their best? *Journal of Policy Analysis and Management, 30*(1), 57-87. doi: 10.1002/pam.20549

Goldhaber, D., & Hansen, M. (2008). *Is it just a bad class? Assessing the stability of measured teacher performance* (Working paper 2010-3). Seattle, WA: Center for Education Data & Research.

Goldhaber, D., Lavery, L., & Theobald, R. (2015). Uneven playing field? Assessing the teacher quality gap between advantaged and disadvantaged students. *Educational Research, 44*(5), 293-307. doi: 10.3102/0013189x15592622

Goldhaber, D., Quince, V., & Theobald, R. (2018). Has it always been this way? Tracing the evolution of teacher quality gaps in U.S. public schools. *American Educational Research Journal, 55*(1), 171-201. doi: 10.3102/0002831217733445

Goldhammer, R. (1969). *Clinical supervision: Special methods for the supervision of teachers*. New York, NY: Holt, Rinehart, & Winston.

Goldrick, L. (2002). *Improving teacher evaluation to improve teaching quality: Issue brief*. Washington, DC: Center for Best Practices, National Governors Association.

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value-added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher, 44*(2), 96-104. doi: 10.301/20013189X15575031

Goldschmidt, P., Choi, K., & Beaudoin, J. B. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance.* Washington, DC: Council of Chief State School Officers.

Goodlad, J. I. (1979). *What schools are for*. Arlington, VA: Phi Delta Kappa Educational Foundation.

Goodlad, J. I. (2003, April 23). A nation in wait. *Education Week, 22*(32), 24-25, 36. Retrieved from http://www.edweek.org/ew/articles/2003/04/23/32goodlad.h22.html

Gould, S. J. (1995). Mismeasure by any measure. In. R Jacoby & N. Glauberman (Eds.), *The bell curve debate: History, documents, opinions* (pp. 3-13). New York, NY: Random House.

Gould, S. J. (1996). *The mismeasure of man*. New York, NY: W. W. Norton Company.

Graves, J. L. (2001). *The emperor's new clothes: Biological theories of race at the millennium*. Piscataway, NJ: Rutgers University Press.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4-27. doi: 10.1037/0033-295x.102.1.4

Grissom, J. A., & Youngs, P. (2016). Making the most of multiple measures. In J. A. Grissom & P. Youngs (Eds.), *Improving teacher evaluation systems: Making the most of multiple measures* (pp. 1-7). New York, NY: Teachers College Press.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher, 43*(6), 293-303. doi: 10.3102/0013189X14544542

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2010). *Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores* (Working Paper 45). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2012). *Can value-added measures of teacher education performance be trusted?* (Working paper #18). East Lansing, MI: The Education Policy Center at Michigan State University.

Gudgel, R. (2011). *Fiscal impact report of SB 502*. Santa Fe, NM: New Mexico Legislature.

Gurung, R., & Vespia, K. (2007). Looking good, teaching well? Linking liking, looks, and learning. *Teaching of Psychology, 34*(1), 5-10. doi: 10.1080/00986280709336641

Guthrie, J. W., & Springer, M. G. (2004). *A Nation at Risk* revisited: Did "wrong" reasoning result in "right" results? At what cost? *Peabody Journal of Education, 79*(1), 7-35. doi: 10.1207/s15327930pje7901_2

Haertel, E., & Herman, J. (2005). *A historical perspective on validity arguments for accountability testing* (CSE Report 654). Los Angeles, CA: Center for the Study of Education, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27. doi: 10.1111/j.1745-3992.2004.tb00149.x

Hamermesch, D. S., & Parker, A. (2005). Beauty in the classroom: Instructor's pulchritude and putative pedagogical productivity. *Economics of Education Review, 24*(4), 369-376. doi: 10.1016/j.econedurev.2004.07.013

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., & Hamagami, A. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *Elementary School Journal, 113*(4), 461-487. doi: 10.1086/669616

Hanushek, E. A. (1970). *The value of teachers in teaching*. Santa Monica, CA: Rand Corporation.

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *The Journal of Human Resources, 14*(3), 351-388. doi: 10.2307/145575

Hanushek, E. A. (2011). Valuing teachers: How much is a good teacher worth? *Education Next, 11*(3), 41-45.

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Economics, 18*(5), 527-544. doi: 10.1002/jae.741

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Why public schools lose teachers. *Journal of Human Resources, 39*(2), 326-354. doi: 10.2307/3559017

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267-271. doi: 10.1257/aer.100.2.267

Harris, D. N. (2011). *Value-added measurement in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press.

Harris, D. N., & Anderson, A. (2013). *Does value-added work better in elementary than in secondary grades?* Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

Harris, D. N., & Herrington, C. D. (2015). Editors' introduction: The use of teacher value-added measures in schools: New evidence, unanswered questions, and future prospects. *Educational Researcher, 44*(2), 71-76. doi: 10.3102/0013189X15576142

Harris, D. N., & Sass, T. R. (2009). *What makes for a good teacher and who can tell?* (Working paper 30). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95*(7-8), 798-812. doi: 10.1016/j.jpubeco.2010.11.009

Harris, W. U. (1981). Teacher command of subject matter. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 58-72). Beverly Hills, CA: SAGE Publications.

Harvey, D. (2007). *A brief history of neoliberalism.* New York, NY: Oxford University Press.

Hazi, H. M., & Rucinski, D. A. (2009). Teacher evaluation as a policy target for improved student learning: A fifty-state review of statute and regulatory action since NCLB. *Education Policy Analysis Archives, 17*(5). doi: 10.14507/epaa.v17n5.2009

Heinz, H. (2011, September 24). N.M. school reform efforts get boost. *Albuquerque Journal*. Retrieved from https://www.abqjournal.com/58725/nm-school-reform-efforts-get-boost.html

Heitin, L. (2012, July 11). Next up in teacher evaluations: Student surveys [Web log post]. *Education Week*. Retrieved from http://blogs.edweek.org/teachers/teaching_now/2012/07/next_up_in_teacher_evaluations_student_surveys.html

Herbert, B. (1994, October 26). In America; Throwing a curve. *New York Times*. Retrieved from https://www.nytimes.com/1994/10/26/opinion/in-america-throwing-a-curve.html

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record, 116*(1).

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York, NY: Simon & Schuster Inc.

Hibler, W. D., & Snyder, J. A. (2015). Teaching matters: Observations on teacher evaluations. *Schools: Studies in Education, 12*(1), 33-47. doi: 10.1086/680693

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*(3), 167-184. doi: 10.1037/h0029780

Hiebert, J., & Morris, A. K. (2012). Teaching, rather than teachers, as a path toward improving classroom instruction. *Journal of Teacher Education, 63*(2), 92-102. doi: 10.1177/0022487111428328

Hill, H. C., Charalambos, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56-64. doi: 10.3102/0013189X12437203

Hill, H. C., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review, 83*(2), 371-384. doi: 10.17763/haer.83.2.d11511403715u376

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831. doi: 10.3102/0002831210387916

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Seattle, WA: Bill & Melinda Gates Foundation.

Hoff, D. J. (1999, March 24). Echoes of the Coleman Report. *Education Week, 18*(28), 3.

Holloway, J. (2019). Teacher evaluation as an onto-epistemic framework. *British Journal of Sociology of Education, 40*(2), 174-189. doi: 10.1080/01425692.2018.1514291

Holloway, J., & Brass, J. (2018). Making accountable teachers: The terrors and pleasures of performativity. *Journal of Education Policy, 33*(3), 361-382. doi: 10.1080/02680939.2017.1372636

Holloway, J., Sørensen, T. B., & Verger, A. (2017). Global perspectives on high-stakes teacher accountability policies: An introduction. *Education Policy Analysis Archives, 25*(85). doi: 10.14507/epaa.25.3325

Holloway-Libell, J. (2015). Evidence of grade and subject-level bias in value-added measures. *Teachers College Record, 117*.

Horng, E. L. (2009). Teacher tradeoffs: Disentangling teachers' preferences for working conditions and student demographics. *American Educational Research Journal, 46*(3), 690-717. doi: 10.3102/0002831208329599

Houston Federation of Teachers (Plaintiff) v. Houston Independent School District (Defendant), Civil No. 4:14-CV-01189. (2015). United States District Court, Southern District of Texas, Houston Division.

Houston Independent School District (HISD). (2012). *HISD Core Initiative I: An effective teacher in every class, teacher appraisal and development system – Year one summary report*. Houston, TX: Author.

Houston Independent School District (HISD). (2013). *Progress conference briefing*. Houston, TX: Author.

Houston, W. R. (1974). Competency based education. In W. R. Houston (Ed.), *Exploring competency based education* (pp. 3-15). Berkeley, CA: McCutchan Publishing Corporation.

Houston, W. R., & Howsam, R. B. (1974). CBTE: The ayes of Texas. *Phi Delta Kappan, 55*(5), 299-303.

Howell, W. G. (2015). Results of President Obama's Race to the Top. *Education Next, 15*(4).

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*(1), 64-86. doi: 10.1037/1082-989x.5.1.64

Hursh, D. (2001). Neoliberalism and the control of teachers, students, and learning: The rise of standards, standardization, and accountability. *Cultural Logic, 4*(1).

Improving America's Schools Act of 1994, Pub. L. No. 103-382 (1994).

Inter-University Consortium for Political and Social Research. (2019). Measures of effective teaching longitudinal database: Applying for access: Frequently asked questions. Retrieved from https://www.icpsr.umich.edu/icpsrweb/content/METLDB/step2/faqs.html

Isenberg, E., Max, J., Gleason, P., Johnson, M., Deutsch, J., & Hansen, M. (2016). *Do low-income students have equal access to effective teachers? Evidence from 26 districts* (NCEE 2017-4008). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Ishii, J., & Rivkin, S. G. (2009). Impediments to the estimation of teacher value added. *Education Finance and Policy, 4*(4), 520-536. doi:10.1162/edfp.2009.4.4.520

Jackson, C. K. (2012). *Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina* (Working Paper No. 18624). Cambridge, MA: National Bureau of Economic Research.

Jackson, C. K. (2014). *Are working conditions related to teacher effectiveness?* Paper presented at the Annual Conference of the Association for Education Finance and Policy (AEFP), San Antonio, TX.

Jackson, C. K. (2019). The full measure of a teacher: Using value-added to assess effects on student behavior. *Education Next, 19*(1), 62-68.

Jacob, B. A., & Lefgren, L. (2007). What do parents value in education? An empirical investigation of parents' revealed preferences for teachers. *Quarterly Journal of Economics, 122*(4), 1603-1637. doi: 10.1162/qjec.2007.122.41603

Jacoby, R., & Glauberman, N. (Eds.) (1995). *The bell curve debate: History, documents, opinions*. New York, NY: Random House.

Jargowsky, P. A., & El Komi, M. (2011). Before or after the bell? School context and neighborhood effects on student achievement. In H. B. Newburger, E. L. Birch, & S. M. Wachter (Eds.), *Neighborhood and life chances: How place matters in modern America* (pp. 50-72). Philadelphia, PA: University of Pennsylvania Press.

Jarrett, H. H. (1977). Implications of implementing competency-based education in the liberal arts. *Educational Technology, 17*(4), 21-26.

Jencks, C., Bartlett, S., Corcoran, M., Crouse, J., Eaglesfield, D., Jackson, G., … Williams, J. (1979). *Who gets ahead? The determinants of economic success in America*. New York, NY: Basic Books.

Jensen, B., Wallace, T. L., Steinberg, M. P., Gabriel, R. E., Dietiker, L., Davis, D. S., … Rui, N. (2019). Complexity and scale in teaching effectiveness research: Reflections from the MET study. *Education Policy Analysis Archives, 27*(7). doi: 10.14507/epaa.27.3923

Jerald, C. D., & Van Hook, K. (2011). *More than measurement: The TAP system's lessons learned for designing better teacher evaluation systems*. Santa Monica, CA: National Institute for Excellence in Teaching.

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York, NY: Guilford Press.

Johnson, S. M., Kraft, M. A., & Papay, J. P. (2012). How context matters in high-needs schools: The effects of teachers' working conditions on their professional satisfaction and their students' achievement. *Teachers College Record, 114*(10).

Jordan-Irvine, J. (1990). *Black students and school failure: Policies, practices, and prescriptions*. New York, NY: Greenwood.

Kalogrides, D., Loeb, S., & Beteille, T. (2013). Systematic sorting: Teacher characteristics and class assignments. *Sociology of Education, 86*(2), 103-123. doi: 10.1177/0038040712456555

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Education Measurement, 38*(4), 319-342. doi: 10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17-64). Washington, D.C.: The National Council on Measurement in Education and American Council on Education.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. doi: 10.1111/jedm.12000

Kane, M. T. (2017). *Measurement error and bias in value-added models* (ETS RR-17-25). Princeton, NJ: Educational Testing Services. doi:10.1002/ets2.12153

Kane, M. T., & Case, S. M. (2014). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17*(3), 221-240. doi: 10.1207/s15324818ame1703_1

Kane, T. J. (2015). Teachers must look in the mirror. *The New York Daily News*. Retrieved from http://www.nydailynews.com/opinion/thomas-kane-teachers-mirror-article-1.2172662

Kane, T. J., & Cantrell, S. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project*. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., Kerr, K. A., & Pianta, R. C. (Eds.). (2014). *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project*. San Francisco, CA: Jossey-Bass.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (Working Paper No. 14607). Cambridge, MA: National Bureau of Economic Research.

Kane, T. J., & Staiger, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., Staiger, D. O., Grissmer, D. & Ladd, H. F. (2002). Volatility in school test scores: Implications for test-based accountability systems. *Brookings Papers on Education Policy, 5*(2002), 235-283. doi: 10.1353/pep.2002.0010

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practice using student achievement data. *Journal of Human Resources, 46*(3), 587-613. doi: 10.3386/w15803

Kantor, H. (1991). Education, social reform, and the state: ESEA and federal education policy in the 1960s. *American Journal of Education, 100*(1), 47-83. doi: 10.1086/444004

Kappler Hewitt, K. (2015). Educator evaluation policy that incorporates EVAAS value-added measures: Undermined intentions and exacerbated inequities. *Education Policy Analysis Archives*, *23*(76), 1-49. doi: 10.14507/epaa.v23.1968

Kauchak, D., Peterson, K. D., & Driscoll, A. (1985). An interview study of teachers' attitudes toward teacher evaluation practices. *Journal of Research and Development in Education, 19*(1), 32-37.

Keith, T. Z. (2015). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling*. New York, NY: Routledge.

Kennedy, M. M. (2010a). Attribution error and the quest for teacher quality. *Educational Researcher, 39*(8), 591-598. doi: 10.3102/0013189X10390804

Kennedy, M. M. (2010b). Introduction: The uncertain relationship between teacher assessment and teacher quality. In M. M. Kennedy (Ed.), *Teacher assessment and the quest for teacher quality: A handbook* (pp. 1-6). San Francisco, CA: Jossey-Bass.

Kersting, N. B., Chen, M., & Stigler, J. W. (2012). Value-added teacher estimates as part of teacher evaluations: Exploring the effects of data and model specifications on the stability of teacher value-added scores. *Education Policy Analysis Archives, 21*(7). doi: 10.14507/epaa.v21n7.2013

Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education, 79*(4), 54-78. doi: 10.1207/s15327930pje7904_4

Kinsler, J. (2012). Assessing Rothstein's critique of teacher value-added models. *Quantitative Economics, 3*(2), 333-362. doi: 10.3982/qe132

Kitto, K., Williams, C., & Alderman, L. (2019). Beyond average: Contemporary statistical techniques for analysing student evaluations of teaching. *Assessment & Evaluation in Higher Education, 44*(3), 338-360. doi: 10.1080/02602938.2018.1506909

Klein, A. (2019, April 2). States, districts tackle the tough work of making ESSA a reality. *Education Week, 38*(27), 4-6. Retrieved from https://www.edweek.org/ew/articles/2019/04/03/states-districts-tackle-the-tough-work-of.html

Knight, D. S. (2019). Are school districts allocating resources equitably? The Every Student Succeeds Act, teacher experience gaps, and equitable resource allocation. *Educational Policy, 33*(4), 615-649. doi: 10.1177/0895904817719523

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function* (Working paper). Columbia, MO: University of Missouri, Department of Economics.

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy, 6*(1), 18–42. doi: 10.1162/EDFP_a_00027

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review, 47*, 180–195. doi: 10.1016/j.econedurev.2015.01.006

Koretz, D. (1992). State and national assessment. In M. C. Alkin (Ed.), *Encyclopedia of Educational Research* (6th ed.) (pp. 1262-1267). Washington, DC: American Educational Research Association.

Koretz, D. (1996). Using student assessments for educational accountability. In E. A. Hanushek & D. W. Jorgensen (Eds.), *Improving America's schools: The role of incentives* (pp. 171-195). Washington, DC: National Academy Press.

Koretz, D. (2017). *The testing charade: Pretending to make schools better*. Chicago, IL: University of Chicago Press.

Kraft, M. A., & Gilmour, A. F. (2016). Can principals promote teacher development as evaluators? A case study of principals' views and experiences. *Educational Administration Quarterly, 52*(5), 711-753. doi: 10.1177/0013161X16653445

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the Widget Effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher, 46*(5), 234-249. doi: 10.3102/0013189X17718797

Kuhfeld, M. (2017). When students grade their teachers: A validity analysis of the Tripod student survey. *Educational Assessment, 22*(4), 253-274. doi: 10.1080/10627197.2017.1381555

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environment Research, 9*(3), 231-251. doi: 10.1007/s10984-006-9015-7

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value-Added Assessment System. *Educational Evaluation and Policy Analysis, 25*(3), 287-298. doi: 10.3102/01623737025003287

Kyriakides, L. (2005). Drawing from teacher effectiveness research and research into teacher interpersonal behaviour to establish a teacher evaluation system: A study on the use of student ratings to evaluate teacher behaviour. *Journal of Classroom Interaction, 40*(2), 44-66.

Ladd, H. F. (2011). Teachers' perceptions of their working conditions: How predictive of planned and actual teacher movement? *Educational Evaluation and Policy Analysis, 33*(1), 235-261. doi: 10.3102/0162373711398128

Ladd, H. F., & Sorensen, L. C. (2017). Returns to teacher experience: Student achievement and motivation in middle school. *Education Finance and Policy, 12*(2), 241-279. doi: 10.1162/edfp_a_00194

LaFee, S. (2014). Students evaluating teachers. *School Administrator, 3*(71), 17-25.

Larkin, D., & Oluwole, J. O. (2014). *The opportunity costs of teacher evaluation: A labor and equity analysis of the TEACHNJ legislation*. New Jersey Educational Policy Forum.

Lavery, M. R., Amrein-Beardsley, A., Geiger, T., & Pivovarova, M. (in press). Value-added model (VAM) scholars on using VAMs for teacher evaluation, post the passage of the Every Student Succeeds Act (ESSA). *Teachers College Record*.

Learning Sciences International. (2017). Marzano teacher evaluation model. Retrieved from http://www.marzanoevaluation.com/evaluation/causal_teacher_evaluation_model/

Lederman (Plaintiff) v. King (Defendant), State of New York, Albany County, Supreme Court (2014).

Lei, X., Li, H., & Leroux, A. (2018). Does a teacher's classroom observation rating vary across multiple classrooms? *Educational Assessment, Evaluation and Accountability, 30*(1), 27-46. doi: 10.1007/s11092-017-9269-x

Leu, E. (2005). *The role of teachers, schools, and communities in quality education: A review of the literature*. Washington, DC: Academy for Educational Development, Global Education Center.

Lewis, S., & Holloway, J. (2018). Datafying the teaching 'profession': Remaking the professional teacher in the image of data. *Cambridge Journal of Education, 49*(1), 35-51. doi: 10.1080/0305764X.2018.1441373

Li, X. (2019, March). *Comparing teachers across subject types and school levels: Evaluating teacher effectiveness with Teacher Effectiveness Student Survey (TESS)*. Paper presented at the Annual Conference of the Association for Education Finance and Policy (AEFP), Kansas City, MO.

Liaw, S. H., & Goh, K. L. (2003). Evidence and control of biases in student evaluations of teaching. *International Journal of Educational Management, 17*(1), 37-43. doi: 10.1108/09513540310456383

Lin, X. (2010). Identifying peer effects in student academic achievement by spatial autoregressive models with group unobservables. *Journal of Labor Economics, 28*(4), 825-860. doi: 10.1086/653506

Lingard, B. (2011). Policy as numbers: Ac/counting for educational research. *Australian Educational Researcher, 38*(4), 355-382. doi: 10.1007/s13384-011-0041-9

Linn, R. L. (2004). *Rethinking the No Child Left Behind accountability system.* Boulder, CO: National Center for Research on Evaluation, Standards, and Student Testing, University of Colorado at Boulder.

Linn, R. L. (2008). Methodological issues in achieving school accountability. *Journal of Curriculum Studies, 40*(6), 699-711. doi: 10.1080/00220270802105729

Linn, R. L., Baker, & Betebenner, D. (2002). *Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001* (CSE Technical Report 567). Los Angeles, CA: Center for the Study of Education, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Linn, R. L., & Haug, C. (2002). *Stability of school building accountability scores and gains* (CSE Technical Report 561). Los Angeles, CA: Center for the Study of Education, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Loeb, S., & Candelaria, C. A. (2012). *How stable are value-added estimates across years, subjects, and student groups?* Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

Loewus, L. (2017, November 4). Are states changing course on teacher evaluation? Test-score growth plays lesser role in six states. *Education Week, 37*(13), 1, 16-17. Retrieved from https://www.edweek.org/ew/articles/2017/11/15/are-states-changing-course-on-teacher-evaluation.html

Lohman, J. (2010). *Comparing No Child Left Behind and Race to the Top* (OLR Research Report 2010-R-0235). Hartford, CT: Connecticut General Assembly.

Lomax, R. G., & Hahs-Vaughn, D. (2012). *Statistical concepts: A second course* (4th ed.). New York, NY: Routledge.

Los Angeles Times. (2010, August 14). Los Angeles teacher ratings. *Los Angeles Times*. Retrieved from http://projects.latimes.com/value-added/

Loup, K., Garland, J., Ellett, C., & Rugutt, J. (1996). Ten years later: Findings from a replication of a study of teacher evaluation practices in our 100 largest school districts. *Journal of Personnel Evaluation in Education, 10*(3), 203-226. doi: 10.1007/bb00124986

Lowrey, A. (2012, January 6). Big study links good teachers to lasting gain. *New York Times*. Retrieved from http://www.nytimes.com/2012/01/06/education/big-study-links-good-teachers-to-lasting-gain.html

Lubienski, C. (2002). Charter schools and privatization. In G. Miron & C. Nelson (Eds.), *What's public about charter school? Lessons learned about choice and accountability* (pp. 1-17). Thousand Oaks, CA: Corwin Press, Inc.

Mansfield, R. (2015). Teacher quality and student inequality. *Journal of Labor Economics, 33*(3), 751-788. doi: 10.1086/679683

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*(3), 253-288. doi: 10.1016/0883-0355(87)90001-2

Marsh, H. W. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology, 83*(2), 285-296. doi: 10.1037//0022-0663.83.2.285

Marsh, H. W. (2007). Students’ evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319-383). New York, NY: Springer.

Marsh, H. W., Dicke, T., & Pfeiffer, M. (2019). A tale of two quests: The (almost) non-overlapping research literatures on students’ evaluations of secondary-school and university teachers. *Contemporary Educational Psychology, 58,* 1-18. doi: 10.1016/j.cedpsych.2019.01.011

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Koller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*(2), 106-124. doi: 10.1080/00461520.2012.670488

Martinez, J. F., Schweig, J., & Goldschmidt, P. (2016). Approaches for combining multiple measures of teacher performance: Reliability, validity, and implications for evaluation policy. *Educational Evaluation and Policy Analysis, 38*(4), 738-756. doi: 10.3102/0162373716666166

Marzano, R. J., Frontier, T., & Livingston, D. (2011). *Effective supervision: Supporting the art and science of teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Mashburn, A. J., Downer, J. T., Rivers, S. E., Brackett, M. A., & Martinez, A. (2014). Improving the power of an efficacy student of a social and emotional learning program: Application of generalizability theory to the measurement of classroom-level outcomes. *Prevention Science, 15*(2), 146-155. doi: 10.1007/s11121-012-0357-3

Mason, B. A., Gunersel, A. B., & Ney, E. A. (2014). Cultural and ethnic bias in teacher ratings of behavior: A criterion-focused review. *Psychology in the Schools, 51*(10), 1017-1030. doi: 10.1002/pits.21800

Mathis, W. J. (2003). No Child Left Behind: Costs and benefits. *Phi Delta Kappan, 84*(9), 679-686.

Maulana, R., & Helms-Lorenz, M. (2016). Observations and student perceptions of the quality of preservice teachers’ teaching behaviour: Construct representation and predictive quality. *Learning Environments Research, 19*(3), 335-357. doi: 10.1007/s10984-016-9215-8

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572–606. doi: 10.1162/edfp.2009.4.4.572

McCoach, D. B., Goldstein, J., Behuniak, P., Reis, S. M., Black, A. C., Sullivan, E. E., & Rambo, K. (2010). Examining the unexpected: Outlier analyses of factors affecting student achievement. *Journal of Advanced Academics, 21*(3), 426-468. doi: 10.1177/1932202x1002100304

McGrady, P. B., & Reynolds, J. R. (2013). Racial mismatch in the classroom: Beyond black-white differences. *Sociology of Education, 86*(1), 3-17. doi: 10.1177/0038040712444857

McGreal, T. L. (1983). *Successful teacher evaluation*. Alexandria, VA: Association for Supervision and Curriculum Development.

McGuinn, P. J. (2006). *No Child Left Behind and the transformation of federal education policy, 1965-2005*. Lawrence, KS: University Press of Kansas.

McGuinn, P. J. (2011). Stimulating reform: Race to the Top, competitive grants, and the Obama education agenda. *Educational Policy, 26*(1), 136-159. doi: 10.1177/0895904811425911

McGuinn, P. J. (2014). Presidential policymaking: Race to the Top, executive power, and the Obama education agenda. *The Forum, 12*(1), 61-79. doi: 10.1515/for-2014-0017

McKown, C., & Weinstein, R. S. (2003). The development and consequences of stereotype consciousness in middle school. *Child Development, 74*(2), 498-515. doi: 10.1111/1467-8624.7402012

McLean, R. L., & Sanders, W. L. (1984). *Objective component of teacher evaluation: A feasibility study* (Working Paper No. 199). Knoxville, TN: University of Tennessee, College of Business Administration.

McNeil, M. (2012, October 17). States punch reset button with NCLB waivers. *Education Week, 32*(8), 1, 25. Retrieved from http://www.edweek.org/ew/articles/2012/10/17/08waiver_ep.h32.html?tkn=TOLFINGBiBfnkhUZdYpn7E0imHiEeu8dd7zM

McPherson, M. A., Todd Jewell, R., & Kim, M. (2009). What determines student evaluation scores? A random effects analysis of undergraduate economics classes. *Eastern Economic Journal, 35*(1), 37-51. doi: 10.1057/palgrave.eej.9050042

Medina, J., & Rich, M. (2016, April 14). California appeals court reverses decision to overturn teacher tenure rules. *New York Times*. Retrieved from https://www.nytimes.com/2016/04/15/us/californiaappealscourt-reverses-decision-to-overturn-teacher-tenure-rules.html

Merrigan, G., & Huston, C. L. (2004). *Communication research methods*. New York, NY: Oxford University Press.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*(10), 955-966. doi: 10.1037//0003-066x.30.10.955

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012-1027. doi: 10.1037//003-066x.35.11.1012

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103.) New York, NY: American Council on Education and Macmillan.

Michelmore, K., & Dynarski, S. (2017). The gap within the gap: Using longitudinal data to understand income differences in student achievement. *AERA Open, 3*(1), 1-18. doi: 10.1177/2332858417692958

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching*. Seattle, WA: Bill & Melinda Gates Foundation.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33-53. doi: 10.1207/S15327930pje7904_3

Morganstein, D., & Wasserstein, R. (2014). ASA statement on value-added models. *Statistics and Public Policy, 1*(1), 108-110. doi: 10.1080/2330443X.2014.956906

Morris, E. W. (2005). From “middle class” to “trailer trash”: Teachers’ perceptions of white students in a predominantly minority school. *Sociology of Education, 78*(2), 99-121. doi: 10.1177/003804070507800201

Mulgrew (Plaintiff) v. Board of Education of City School District of New York (Defendant), NY Slip Op 21252. (2011). Supreme Court, New York County.

Nadeem, M. (2013, December 22). Teachers wary of new evaluation system in New Mexico. *Education News*. Retrieved from http://www.educationnews.org/education-policy-and-politics/teachers-wary-of-new-evaluation-system-in-new-mexico/

Nagel, J. (1994). Constructing ethnicity: Creating and recreating ethnic identity and culture. *Social Problems, 41*(1), 152-176. doi: 10.1525/sp.1994.41.1.03x0430n

Nasir, N. S., & Hand, V. M. (2006). Exploring sociocultural perspectives on race, culture, and learning. *Review of Educational Research, 76*(4), 449-475. doi: 10.3102/00346543076004449

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: United States Government Printing Office. Retrieved from http://www2.ed.gov/pubs/NatAtRisk/index.html

National Council on Teacher Quality. (2015). *2015 state teacher policy yearbook: New Mexico*. Washington, DC: Author.

National Council on Teacher Quality. (2017). *Running in place: How new teacher evaluations fail to live up to the promises*. Washington, DC: Author.

National Education Association New Mexico. (2018). Common questions about us. Retrieved from https://nea-nm.org/questions-about-us/

National Institute for Excellence in Teaching. (2017). Elements of success. Retrieved from http://www.niet.org/tap-system/elements-of-success/

Naureckas, J. (1995, January 1). Racism resurgent: How media let *The Bell Curve's* pseudo-science define the agenda on race [Web log post]. *FAIR.org*. Retrieved from https://fair.org/home/racism-resurgent/

Nee, E. (2010). Q & A: Joanne Weiss. *Stanford Social Innovation Review, 8*(2), 13-15.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Cici, S. J., … Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychology, 51*(2), 77-101.

New Mexico Administrative Code (NMAC). (2011). Performance Evaluation System Requirements for Teachers. 6.69.4

New Mexico Administrative Code (NMAC). (2012). Primary and Secondary Education, School Personnel – Performance, Teacher and School Leader Effectiveness. 6.69.8.

New Mexico Administrative Code (NMAC). (2017). Primary and Secondary Education, School Personnel – Performance, Teacher and School Leader Effectiveness. 6.69.8.

New Mexico Alliance for Minority Participation. (2016). Undergraduate research scholars (URS). Retrieved from https://nmamp.nmsu.edu/undergraduate-research-scholars-urs/

New Mexico Effective Teaching Task Force. (2011). *Final report and recommendations*. Santa Fe, NM: Author.

New Mexico Legislature. (2003). *Revised summary of selected provisions: HB 212* Public School Reforms. Santa Fe, NM: Author.

New Mexico Public Education Department. (n.d.a). *New Mexico rising: An executive summary of New Mexico's state plan for the Every Student Succeeds Act*. Santa Fe, NM: Author.

New Mexico Public Education Department. (n.d.b). *NMTEACH Domain 1: Planning and Preparation*. Santa Fe, NM: Author.

New Mexico Public Education Department. (n.d.c). *NMTEACH Domain 2: Creating an Environment for Learning*. Santa Fe, NM: Author.

New Mexico Public Education Department. (n.d.d). *NMTEACH Domain 3: Teaching for Learning*. Santa Fe, NM: Author.

New Mexico Public Education Department. (n.d.e). *NMTEACH Domain 4: Professionalism*. Santa Fe, NM: Author.

New Mexico Public Education Department. (n.d.f). *Parent survey with crosswalk to NMTEACH rubric*. Santa Fe, NM: Author.

New Mexico Public Education Department. (n.d.g). *Student survey with crosswalk to NMTEACH rubric*. Santa Fe, NM: Author.

New Mexico Public Education Department. (2010a). *New Mexico's Race to the Top application for initial funding*. Sante Fe, NM: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/phase1-applications/new-mexico.pdf

New Mexico Public Education Department. (2010b). *New Mexico's Race to the Top application for initial funding*. Sante Fe, NM: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/phase2-applications/new-mexico.pdf

New Mexico Public Education Department. (2011). *Governor Susana Martinez highlights education reform agenda in first state of the state address*. Santa Fe, NM: Author.

New Mexico Public Education Department. (2016a). *New Mexico rising: Engaging our communities for excellence in education*. Albuquerque, NM: New Mexico First.

New Mexico Public Education Department. (2016b). *NMTEACH technical guide. Business rules and calculations. 2015-2016*. Santa Fe, NM: Author.

New Mexico Public Education Department. (2018a). *New Mexico rising, together: Building on a foundation of success* [PPT slides]. Santa Fe, NM: Author.

New Mexico Public Education Department. (2018b). *NMTEACH Domains 2 & 3 – Observations*. Santa FE, NM: New Mexico Public Education Department.

New Mexico Public Education Department. (2019a). Achievement data. Retrieved from https://webnew.ped.state.nm.us/bureaus/accountability/achievement-data/

New Mexico Public Education Department. (2019b). *Teacher evaluation SY2018-19*. Santa Fe, NM: Author. Retrieved from https://webnew.ped.state.nm.us/wp-content/uploads/2019/09/Explanation-of-SY2018-19-Evaluation-Report-Design_FINAL.pdf

New Mexico Public Education Department. (2019c). *What's new in teacher evaluation?* Santa Fe, NM: Author. Retrieved from https://webnew.ped.state.nm.us/wp-content/uploads/2019/09/One-Pager-Changes-to-Teacher-Evaluation.pdf

New Mexico Statute. (2011). § 22-1-4: Public schools article: General provisions: Free public schools; exceptions; withdrawing and enrolling; open enrollment. (1996 through 1st Session 50th Legislation).

New York State United Teachers Association (Plaintiff) v. Board of Regents of the University of the State of New York (Defendant), 929 NYS3d 699. (2011). Supreme Court, Albany County.

Newby, R. G., & Newby, D. E. (1994). *The bell curve*: Another chapter in the continuing political economy of racism. *American Behavioral Scientist, 39*(1), 12-24. doi: 10.1177/0002764295039001003

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18*(23). doi: 10.14507/epaa.v18n23.2010

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115 Stat. 1425 (2002).

Noe, J. N., Tocci, C. M., Holtzman, S. L., & Williams, J. C. (2013). *Foundations of observations: Considerations for developing a classroom observation system that helps districts achieve consistent and accurate scores.* Princeton, NJ: Educational Testing Services and Seattle, WA: Bill & Melinda Gates Foundation.

Nott, R. (2014, October 14). Lawmakers hear school leaders' concerns about teacher eval system. *Santa Fe New Mexican*. Retrieved from http://www.santafenewmexican.com/news/education/lawmakers-hear-school-leaders-concerns-about-teacher-eval-system/article_738fa633-11f6-598a-b4bc-f15c86f08e9c.html

O'Donnell, P. (2014, May 11). Ohio students soon could be grading their own teachers. *The Plain Dealer*. Retrieved from http://www.cleveland.com/metro/index.ssf/2014/05/ohio_students_could_soon_be_gr.html

Office of the Governor. (2011). New Mexico Effective Teacher Task Force. Retrieved from http://www.governor.state.nm.us/Effective_Teaching_Task_Force.aspx

Okonofua, J. A., & Eberhardt, J. L. (2015). Two strikes: Race and the disciplining of young students. *Psychological Science, 26*(5), 617-624. doi: 10.1177/0956797615570365

Olson, L. (2004a, August 11). Critics float 'No Child' revisions. *Education Week, 23*(44), 1, 33. Retrieved from http://www.edweek.org/ew/articles/2004/08/11/44alter.h23.html

Olson, L. (2004b, December 8). Taking root. *Education Week*, *24*(15), S1, S3, S7. Retrieved from http://www.edweek.org/ew/articles/2004/12/08/15nclb-1.h24.html

Orland, M. (2015). Research and policy perspectives on data-based decision making in education. *Teachers College Record, 117*(4).

Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation, 8*(2).

Packer, J. (2004, July). *No Child Left Behind and Adequate Yearly Progress fundamental flaws: A forecast for failure*. Paper presented at the Center on Education Policy forum on ideas to improve the accountability provisions under the No Child Left Behind Act, Washington, DC.

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193. doi: 10.3102/0002831210362589

Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review, 82*(1), 123-141. doi: 10.17763/haer.82.1.v40p0833345w6384

Pathe, S., & Choe, J. (2013, February 4). A brief overview of teacher evaluation controversies. *PBS Newshour*. Retrieved from http://www.pbs.org/newshour/rundown/teacher-evaluation-controversies/

Paufler, N. A., & Amrein-Beardsley, A. (2014). The random assignment of students into elementary classrooms: Implications for value-added analyses and interpretations. *American Educational Research Journal, 51*(2), 328-362. doi: 10.3102/0002831213508299

Paul, L. (2015, October 30). NMTEACH 101 for governing council members [Powerpoint presentation]. Santa Fe, NM: New Mexico Public Education Department.

Pecheone, R., & Wei, R. C. (2009). *Review: The widget effect: Our national failure to acknowledge and act on teacher differences*. Boulder, CO: National Education Policy Center.

Pelham, B. (2013). *Intermediate statistics: A conceptual course*. Thousand Oaks, CA: SAGE.

Peterson, E. R., Rubie-Davies, C., Osborne, D., & Sibley, C. (2016). Teachers' explicit expectations and implicit prejudiced attitudes to educational achievement: Relations with student achievement and the ethnic achievement gap. *Learning and Instruction, 42*, 123-140. doi: 10.1016/j.learninstruc.2016.01.010

Peterson, K. D. (2004). Research on school teacher evaluation. *NASSP Bulletin, 88*(639), 60-79. doi: 10.1177/019263650408863906

Peterson, K. D., Stevens, D., & Ponzio, R. (1997). Variable data sources in teacher evaluation. *Journal of Research and Development in Education, 31*(3), 123-132.

Peterson, K. D., Wahlquist, C., & Bone, K. (2000). Student surveys for school teacher evaluation. *Journal of Personnel Evaluation in Education, 14*(2), 135-153.

Peterson, P. E. (2003). *Our schools and our future: Are we still at risk?* Stanford, CA: Hoover Institution Press.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processed: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109-119. doi: 10.3102/0013189X09332374

Pike, G. R. (1999). The constant error of the halo in educational outcomes research. *Research in Higher Education 40*(1), 61-86.

Piper, M. K., & Houston, R. W. (1980). The search for teacher competence: CBTE and MCT. *Journal of Teacher Education, 31*(5), 37-40. doi: 10.1177/002248718003100510

Pivovarova, M., Broatch, J., & Amrein-Beardsley, A. (2014). Chetty et al. on the American Statistical Association's recent position statement on value-added models (VAMs): Five points of contention [Commentary]. *Teachers College Record*. Retrieved from https://www.tcrecord.org/content.asp?contentid=17633

Podgursky, M., & Ballou, D. (2001). *Personnel policy in charter schools*. Washington, DC: The Thomas B. Fordham Institute.

Podolsky, A., Kini, T., & Darling-Hammond, L. (2019). Does teaching experience increase teacher effectiveness? A review of US research. *Journal of Professional Capital and Community, 4*(4), 286-308. doi: 10.1108/jpcc-12-2018-0032

Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education, 121*(2), 183-212. doi: 10.1086/679390

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis, 36*(4), 399-416. doi: 10.3102/0162373714531851

Poon, A., & Schwartz, N. (2016, March 17). *Investigating misalignment in teacher observation and value-added ratings*. Paper presented at the annual meeting of the Association for Education Finance and Policy, Denver, CO.

Poovey, M. (1998). *A history of the modern fact: Problems of knowledge in the sciences of wealth and society*. Chicago, IL: University of Chicago Press.

Popham, W. J. (2013). *Evaluating America's teachers: Mission possible?* Thousand Oaks, CA: Corwin Press.

Porter, A. C., Youngs, P., & Odden, A. (2001). Advances in teacher assessment and their uses. In V. Richardson (Ed.), *Handbook of research on teaching* (pp. 259-297). New York, NY: Macmillan.

Porter, T. (1995). *Trust in numbers: The invention of objectivity*. Princeton, NJ: Princeton University Press.

Pugach, M. C., & Raths, J. D. (1983). Testing teachers: Analysis and recommendations. *Journal of Teacher Education, 34*(1), 37-43. doi: 10.1177/002248718303400109

Pullin, D. (2013). Legal issues in the use of student test scores and value-added model (VAM) to determine educational quality. *Education Policy Analysis Archives, 21*(6). doi: 10.14507/epaa.v21n6.2013

Putnam, H., Ross, E., & Walsh, K. (2018). *Making a difference: Six places where teacher evaluation systems are getting results*. Washington, DC: National Council on Teacher Quality.

Quillian, L. (2014). Does segregation create winners and losers? Residential segregation and inequality in educational attainment. *Social Problems, 61*(1), 402-426. doi: 10.1525/sp.2014.12193

Quinn, D. M., & Stewart, A. M. (2019). Examining the racial attitudes of white Pre-K—12 educators. *The Elementary School Journal, 120*(2), 272-299. doi: 10.1086/705899

Quirk, T. J. (1974). Some measurement issues. In W. R. Houston (Ed.), *Exploring competency based education* (pp. 251-260). Berkeley, CA: MrCutrhan Publishing Corporation.

Raudenbush, S. W. & Jean, M. (2012). *How should educators interpret value-added scores?* Stanford, CA: Carnegie Knowledge Network.

Raudenbush, S. W., & Jean, M. (2014). To what extent do student perceptions of classroom quality predict teacher value added? In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project* (pp. 170-202). San Francisco, CA: Jossey-Bass.

Ravitch, D. (2014, August 11). The holes in the Chetty et al VAM study as seen by the American Statistical Association [Web log post]. Retrieved from https://dianeravitch.net/2014/08/11/the-holes-in-the-chetty-et-al-vam-study-as-seen-by-the-american-statistical-association/comment-page-1/

Rebell, M. A., & Hunter, M. A. (2004). ‘Highly qualified’ teachers: Pretense or legal requirement? *The Phi Delta Kappan, 85*(9), 690-696.

Reiss, R. (2017). A vindication of the criticism of New Mexico Public Education Department’s teacher evaluation system. *The Beacon, XX*(1), 2-4. Retrieved from http://www.cese.org/wp-content/uploads/2017/05/2017-05-Beacon.pdf

Remedios, R., & Lieberman, D. A. (2008). I liked your course because you taught me well: The influence of grades, workload, expectations and goals on students’ evaluations of teaching. *British Educational Research Journal, 34*(1), 91-115. doi: 10.1080/01411920701492043

Rentner, D. S., Chudowsky, N., Fagan, T., Gayler, K., Hamilton, M., & Kober, N. (2003). *From the capital to the classroom: State and federal efforts to implement the No Child Left Behind Act*. Washington, DC: Center on Education Policy.

Riley, R. W. (1995). *The Improving America’s Schools Act of 1994: Reauthorization of the Elementary and Secondary Education Act*. Washington, DC: U.S. Department of Education.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458. doi: 10.1111/j.1468-0262.2005.00584.x

Rivkin, S. G., & Ishii, J. (2008). *Impediments to the estimation of teacher value-added*. Paper presented at the National Conference on Value-Added Modeling (VAM), Madison, WI.

Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review, 94*(2), 247-252. doi: 10.1257/0002828041302244

Rose, N. (1991). Governing by numbers: Figuring out democracy. *Accounting, Organizations and Society, 16*(7), 673-692. doi: 10.1016/0361-3682(91)90019-b

Rose, N. (1999). *The powers of freedom: Reframing political thought*. New York, NY: Cambridge University Press.

Rosenbaum, P., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55. doi: 10.2307/2335942

Ross, E., Gerber, N., Jarmolowski, H., Lakis, K., Ledyard, N., Staresina, L., & Worth, C. (2017). *2017 state teacher policy yearbook: National summary*. Washington, DC: National Council on Teacher Quality.

Ross, E., & Walsh, K. (2019). *State of the states 2019: Teacher and principal evaluation policy*. Washington, DC: National Council on Teacher Quality.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10) (pp. 173-220). New York, NY: Academic Press.

Ross, S. M., Smith, L. J., Alberg, M., & Lowther, D. (2004). Using classroom observation as a research and formative evaluation tool in educational reform: The school observation measure. In H. C. Waxman, R. G. Tharp, & R. S. Hilberg (Eds.), *Observational research in U.S. classrooms: New approaches for understanding cultural and linguistic diversity* (pp. 144-173). Cambridge, UK: Cambridge University Press.

Roth, R. A. (1977). How effective are CBTE programs? *Phi Delta Kappa, 58*(10), 757-760.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*(4), 537-571. doi: 10.1162/edfp.2009.4.4.537

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics, 125*(1), 175-214. doi: 10.1162/qjec.2010.125.1.175

Rothstein, J. (2015). *Revisiting the impacts of teachers* (Working paper). Retrieved from http://eml.berkeley.edu/~jrothst/workingpapers/rothstein_cfr.pdf

Rothstein, J. (2017). *Revisiting the impacts of teachers* (Working paper). Berkeley, CA: University of California, Berkeley.

Rothstein, J., & Mathis, W. J. (2013). *Review of two culminating reports from the MET Project*. Boulder, CO: National Education Policy Center.

Rowley, J. F. S., Phillips, S. F., & Ferguson, R. F. (2019). The stability of student ratings of teacher instructional practice: Examining the one-year stability of the 7Cs composite. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice, 30*(4), 549-562. doi: 10.1080/09243453.2019.1620293

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*(3), 299-311. doi: 10.1007/bf00973726

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247-256.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Assessment System: A quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

Sanders, W. L., Wright, S. P., Rivers, J. C., & Leandro, J. G. (2009). *A response to criticisms of SAS EVAAS*. Cary, NC: SAS Institute Inc.

Sandilos, L. E., Sims, W. A., Norwalk, K. E., & Reddy, L. A. (2019). Converging on quality: Examining multiple measures of teaching effectiveness. *Journal of School Psychology, 74*, 10-28. doi: 10.1016/j.jsp.2019.05.004

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation*. Chicago, IL: Consortium on Chicago School Research.

SAS Institute, Inc. (2019). SAS EVAAS for K-12. Retrieved from http://www.sas.com/en_us/industry/k-12-education/evaas.html

Sass, T. R. (2008). *The stability of value-added measures of teacher quality and implications for teacher compensation policy*. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D. N., & Feng, L. (2012). Value added of teachers in high-poverty schools and lower poverty schools. *Journal of Urban Economics, 72*(2-3), 104-122. doi: 10.1016/j.jue.2012.04.004

Sawchuk, S. (2015, October 6). Teacher evaluation heads to the courts. *Education Week, 35*(7), 15. Retrieved from http://www.edweek.org/ew/section/multimedia/teacher-evaluation-heads-to-the-courts.html

Scherrer, J. (2011). Measuring teaching using value-added modeling: The imperfect panacea. *NASSP Bulletin, 95*(20, 122-140. doi: 10.1177/0192636511410052

Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics, 38*(2), 142-171. doi: 10.3102/1076998611432174

School Personnel Act. New Mex. Stat. § 22-10A-19 (2006).

School Personnel Evaluation System. S.B. 503, 50th Leg. (1st session). (2011).

School Teacher and Principal Evaluation. S.B. 502, 50th Leg. (1st session). (2011).

Schulz, J., Sud, G., Crowe, B. (2014). *Lessons from the field: The role of student surveys in teacher evaluation and development*. Sudbury, MA: Bellweather Education Partners.

Schwartz, R. B., & Robinson, M. A. (2000). Goals 2000 and the standards movement. *Brookings Papers on Education Policy, 2000*(1), 173-206. doi: 10.1353/pep.2000.0016

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis, 36*(3), 259-280. doi: 10.3102/0162373713509880

Shaw, L. H. & Bovaird, J. A. (2011, April). *The impact of latent variable outcomes on value-added models of intervention efficacy*. Paper presented at the Annual Conference of the American Educational Research Association (AERA), New Orleans, LA.

Shober, A. F., Manna, P., & Witte, J. F. (2006). Flexibility meets accountability: State charter school laws and their influence on the formation of charter schools in the United States. *Policy Studies Journal, 34*(4), 563-587. doi: 10.1111/j.1541-0072.2006.00191.x

Simpson, R. L., LaCava, P. G., Graner, P. S. (2004). The No Child Left Behind Act: Challenges and implications for educators. *Intervention in School and Clinic, 40*(2), 67-76. doi: 10.1177/10534512040400020101

Skourdoumbis, A., & Gale, T. (2013). Classroom teacher effectiveness research: A conceptual critique. *British Educational Research Journal, 39*(5), 892-906. doi: 10.1002/berj.3008

Smith, S. W., Yoo, J. H., Farr, A. C., Salmon, C. T., & Miller, V. D. (2007). The influence of student sex and instructor sex on student ratings of instructors: Results from a college of communication. *Women's Studies in Communication, 30*(1), 64-77. doi: 10.1080/07491409.2007.10162505

Smolkowski, K., & Gunn, B. (2012). Reliability and validity of the classroom observations of student-teacher interactions (COSTI) for kindergarten reading instruction. *Early Childhood Research Quarterly, 27*(2), 316-328. doi: 10.1016/j.ecresq.2011.09.004

Snyder, T. D., de Brey, C., & Dillow, S. A. (2019). *Digest of education statistics, 2018* (NCES 2020-009). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Soar, R. S. (1973). Accountability: Assessment problems and possibilities. *Journal of Teacher Education*, *24*(3), 205-212. doi: 10.1177/002248717302400307

Song, J. (2012, October 18). Times sues L.A. Unified for teacher ratings [Web log post]. *Los Angeles Times*. Retrieved from http://latimesblogs.latimes.com/lanow/2012/10/times-sues-la-unified-for-teacher-ratings.html

Spooren, P. (2010). On the credibility of the judge: A cross-classified multilevel analysis on student evaluations of teaching. *Studies in Educational Evaluation, 36*(4), 121-131. doi: 10.1016/j.stueduc.2011.02.001

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research, 83*(4), 598-642. doi: 10.3102/0034654313496870

Starr, P. (1987). The sociology of official statistics. In W. Alonso & P. Starr (Eds.), *The politics of numbers* (pp. 7-58). New York, NY: Russell Sage Foundation.

StataCorp. (n.d.). Stata 14. Retrieved from https://www.stata.com/stata14/

State of New Mexico. (2011). Executive Order 2011-024: Formation of the New Mexico Effective Teacher Task Force. Santa Fe, NM: Author.

State of New Mexico. (2016). Executive Order 2016-037: “Route to 66” as New Mexico’s higher education attainment goal. Santa Fe, NM: Author.

State of New Mexico. (2019). Executive Order 2019-002: Directive to the state Public Education Department to immediately take the steps necessary to begin transitioning away from use of the standardized test terms the “Partnership for Assessment of Readiness for College and Careers” (“PARCC”) and to work with stakeholders to identify and implement a more effective method for assessing teacher performance. Santa Fe, NM: Author.

State of New Mexico Ex Rel., The Honorable Mimi Stewart, National Education Association-New Mexico, Manessa Young Padilla, and Deborah Romeo (Plaintiffs) v. Education Secretary-Designee Hanna Skandera in her official capacity of the New Mexico Public Education Department (Defendant); State of New Mexico, County of Santa Fe, First Judicial District Court (2014).

State of New Mexico Ex Rel., The Honorable Mimi Stewart, The Honorable Sheryl Williams Stapleton, The Honorable Howie C. Morales, The Honorable Linda M. Lopez, The Honorable William P. Soules, American Federation of Teachers – New Mexico, Albuquerque Federation of Teachers, Jolene Begay, Dana Allen, Naomi Daniel, Ron Lavondoski, Tracey Brumlik, Crystal Herrera, and Allison Hawks (Plaintiffs) v. New Mexico Public Education Department and Secretary-Designee Hanna Skandera in her official capacity (Defendants); State of New Mexico, County of Bernanillo, First Judicial District Court (2015).

State of New Mexico Ex Rel., The Honorable Sheryl Williams Stapleton, The Honorable Howie C. Morales, The Honorable Linda M. Lopez, American Federation of Teachers – New Mexico, Ellen Bernstein, and Ryan Ross (Plaintiffs) v. Hanna Skandera, Secretary-Designate of the Public Education Department of the State of New (Defendants); State of New Mexico, County of Bernanillo, First Judicial District Court (2013).

Stecher, B., Holtzman, D. J., Garet, M. S., Hamilton, L. S. Engberg, J., Steiner, E. D., … Chambers, J. (2018). *Improving teaching effectiveness: Final report*. Santa Monica, CA: RAND Corporation.

Stedman, J. B. (1994a). *Goals 2000: Overview and analysis* (CRS-94-490-EPW). Washington, DC: Library of Congress, Congressional Research Service.

Stedman, J. B. (1994b). *Improving America's Schools Act: An overview of P.L. 103-382* (CRS-94-872-EPW). Washington, DC: Library of Congress, Congressional Research Service.

Steele, J. L., Pepper, M. J., Springer, M. G., & Lockwood, J. R. (2015). The distribution and mobility of effective teachers: Evidence from a large, urban school district. *Economics of Education Review, 48*(5), 86-101. doi: 10.1016/j.econedurev.2015.05.009

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy, 11*(3), 340-359. doi: 10.1162/edfp_a_00186

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis, 38*(2), 293-317. doi: 10.3102/0162373715616249

Steinberg, M. P., & Sartain, L. (2015). Does teaching evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy, 10*(4), 535-572. doi: 10.1162/EDFP_a_00173

Stevens, T., Harris, G., Liu, X., & Aguirre-Munoz, Z. (2013). Students' ratings of teacher practices. *International Journal of Mathematical Education in Science and Technology, 44*(7), 984-995. doi: 10.1080/0020739X.2013.823250

Stock, J. H., & Watson, M. W. (2007). *Introduction to econometrics* (2nd ed.). Boston, MA: Addison-Wesley.

Stronge, J. H., & Ostrander, L. (1997). Client surveys in teacher evaluation. In J. H. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practice* (pp. 129-161). Thousand Oaks, CA: Corwin Press.

Stronge, J. H., & Tucker, P. D. (2003). *Teacher evaluation: Assessing and improving performance*. Larchmont, NY: Eye on Education.

Swedien, J. (2015, March 31). Feds extend NM No Child waiver for four years. *Albuquerque Journal*. Retrieved from https://www.abqjournal.com/562869/feds-extend-nm-no-child-waiver-for-four-years.html

Sweetnam, A. (1996). The changing contexts of gender: Between fixed and fluid experience. *Psychoanalytic Dialogues, 6*(4), 437-459. doi: 10.1080/10481889609539130

Taubman, P. M. (2009). *Teaching by numbers: Deconstructing the discourse of standards and accountability in education*. New York, NY: Routledge.

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review, 102*(7), 3628-3651. doi: 10.1257/aer.102.7.3628

Taylor, J., Stecher, B., O'Day, J., Naftel, S., Le Floch, K. C. (2010). *State and local implementation of the No Child Left Behind Act Volume IX—Accountability under NCLB: Final report*. Washington, DC: U.S. Department of Education.

Teacher and School Leader Effectiveness Act. H.B. 249, 50th Leg. (2nd session). (2012).

Teacher Choice Compensation Fund. S.B. 567, 50th Leg. (1st session). (2011).

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., … Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*(1), 11-36. doi: 10.3102/10769986029001011

Tennessee State Board of Education. (2012). *Teacher and principal evaluation policy*. Nashville, TN: Author. Retrieved from https://www.tn.gov/assets/entities/sbe/attachments/7-27-12-II_C_Teacher_and_Principal_Evaluation_Revised.pdf

The Danielson Group. (2013). General questions about the framework. Retrieved from http://www.danielsongroup.org/questions-about-the-framework-for-teaching/

The Public School Reform Bill. H.B. 212, 46th Leg. (1st session). (2003).

The New Teacher Project. (2011). *Rating a teacher observation tool: Five ways to ensure classroom observations are focused and rigorous*. Brooklyn, NY: Author.

The White House. (2009). Fact sheet: The race to the top. Washington, DC: Author. Retrieved from https://www.whitehouse.gov/the-press-office/fact-sheet-race-top

The White House. (2012). Remarks by the President in State of the Union address. Retrieved from https://obamawhitehouse.archives.gov/the-press-office/2012/01/24/remarks-president-state-union-address

Thomas, N. C. (1983). The development of federal activism in education: A contemporary perspective. *Education and Urban Society, 15*(3), 271-290. doi: 10.1177/0013124583015003002

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25-29. doi: 10.1037/h0071663

Tobiason, G. (2018). Talking our way around expert caution: A rhetorical analysis of VAM. *Educational Researcher, 48*(1), 19-30. doi: 10.3102/0013189X18797618

Tripod Education Partners. (2017). Learn about Tripod. Retrieved from http://tripoded.com/about-us-2/

Tucker, P. D., & Stronge, J. H. (2005). *Linking teacher evaluation and student learning*. Alexandria, VA: Association for Supervision and Curriculum Development.

U.S. Department of Education. (n.d.a). *Race to the Top: Phase 1 final results*. Washington, DC: Author.

U.S. Department of Education. (n.d.b). *Race to the Top: Phase 2 final results*. Washington, DC: Author.

U.S. Department of Education. (2003). *No Child Left Behind: A parents guide*. Washington, DC: Author.

U.S. Department of Education. (2004). *No Child Left Behind: A toolkit for teachers*. Washington, DC: Author.

U.S. Department of Education. (2005a). Important aspects of No Child Left Behind relevant to NAEP. Retrieved from http://nces.ed.gov/nationsreportcard/nclb.aspx

U.S. Department of Education. (2005b). New No Child Left Behind flexibility: Highly qualified teachers. Retrieved from http://www2.ed.gov/nclb/methods/teachers/hqtflexibility.html

U.S. Department of Education. (2005c). Part A - Improving basic programs operated by local education agencies. Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/pg2.html#sec1111

U.S. Department of Education. (2008a). Growth models: Ensuring grade-level proficiency for all students by 2014. Retrieved from https://www2.ed.gov/admins/lead/account/growthmodel/proficiency.html

U.S. Department of Education. (2008b). *Mapping New Mexico's educational progress 2008*. Washington, DC: Author. Retrieved from https://www2.ed.gov/nclb/accountability/results/progress/newmexico.pdf

U.S. Department of Education. (2009a). *Race to the Top Program executive summary*. Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

U.S. Department of Education. (2009b). Secretary Spellings approves additional growth model pilots for 2008-2009 school year. Retrieved from https://www2.ed.gov/news/pressreleases/2009/01/01082009a.html

U.S. Department of Education. (2010a). 16 finalists announced in Phase I of Race to the Top competition; Finalists to present in mid-March; Winners announced in early April. Retrieved from https://www2.ed.gov/news/pressreleases/2010/03/03042010.html

U.S. Department of Education. (2010b). Delaware and Tennessee win first Race to the Top grants. Retrieved from https://www.ed.gov/news/press-releases/delaware-and-tennessee-win-first-race-top-grants

U.S. Department of Education. (2010c). Nine states and the District of Columbia win second round Race to the Top grants. Retrieved from https://www.ed.gov/news/press-releases/nine-states-and-district-columbia-win-second-round-race-top-grants

U.S. Department of Education. (2010d). *Race to the Top: Panel review by applicant for New Mexico, phase 1*. Washington, DC: Author.

U.S. Department of Education. (2010e). *Race to the Top: Panel review by applicant for New Mexico, phase 2*. Washington, DC: Author.

U.S. Department of Education. (2010f). *Race to the Top: Technical review form – Tier 1. New Mexico application #3600NM-10*. Washington, DC: Author.

U.S. Department of Education. (2010g). *Race to the Top: Technical review form – Tier 1. New Mexico application #4680NM-1*. Washington, DC: Author.

U.S. Department of Education. (2010h). *Race to the Top program: Guidance and frequently asked questions*. Washington, DC: Author.

U.S. Department of Education. (2011a). Department of Education awards $200 million to seven states to advance K-12 reform. Retrieved from https://www.ed.gov/news/press-releases/department-education-awards-200-million-seven-states-advance-k-12-reform

U.S. Department of Education. (2011b). *Letter to Hanna Skandera regarding New Mexico's ESEA flexibility request*. Washington, DC: Author.

U.S. Department of Education. (2011c). *New Mexico – ESEA flexibility request – November 14, 2011*. Washington, DC: Author.

U.S. Department of Education. (2011d). *Race to the Top Phase 3 application overview*. Washington, DC: Author.

U.S. Department of Education. (2012a). Department of Education approves New Mexico's request for flexibility from No Child Left Behind. Retrieved from https://www.ed.gov/news/press-releases/department-education-approves-new-mexicos-request-flexibility-no-child-left-behi

U.S. Department of Education. (2012b). *ESEA flexibility: Frequently asked questions*. Washington, DC: Author. Retrieved from http://www2.ed.gov/policy/eseaflex/esea-flexibility-faqs.doc

U.S. Department of Education. (2012c). *New Mexico ESEA flexibility request – February 15, 2012*. Washington, DC: Author.

U.S. Department of Education. (2012d). *New Mexico ESEA flexibility request – November 9, 2012*. Washington, DC: Author.

U.S. Department of Education. (2014a). *ESEA flexibility: Guidance for renewal process*. Washington, DC: Author.

U.S. Department of Education. (2014b). ESEA flexibility one-year extension. Washington, DC: Author. Retrieved from https://www2.ed.gov/policy/elsec/guid/esea-flexibility/extension/index.html

U.S. Department of Education. (2014c). *Letter to Hanna Skandera regarding New Mexico's ESEA flexibility one-year extension*. Washington, DC: Author.

U.S. Department of Education. (2015a). *Letter to Hanna Skandera regarding New Mexico's request for renewal of flexibility*. Washington, DC: Author.

U.S. Department of Education. (2015b). New Mexico highly qualified teacher letter. Washington, DC: Author. Retrieved from https://www2.ed.gov/policy/eseaflex/secretary-letters/nm4hqtltr.html

U.S. Department of Education. (2015c). U.S. Department of Education approves ESEA flexibility renewal for five states through expedited process. Retrieved from https://www.ed.gov/news/press-releases/us-department-education-approves-esea-flexibility-renewal-five-states-through-expedited-decision-process

U.S. Department of Education. (2016a). *ESSA accountability chart*. Washington, DC: Author. Retrieved from https://ed.gov/policy/elsec/leg/essa/essacctchart1127.pdf

U.S. Department of Education. (2016b). *Every Student Succeeds Act. Accountability, state plans, and data reporting: Summary of final regulations*. Washington, DC: Author.

U.S. Department of Education. (2017a). *Accountability under Title I, Part A of the ESEA: Frequently asked questions*. Washington, DC: Author. Retrieved from https://www2.ed.gov/programs/titleiparta/eseatitleiaccountabilityfaqs.pdf

U.S. Department of Education. (2017b). ESSA state plan submission. Retrieved from https://www2.ed.gov/admins/lead/account/stateplan17/statesubmission.html

U.S. Department of Education. (2017c). Letter to New Mexico from Secretary DeVos regarding state plan. Retrieved from https://www2.ed.gov/admins/lead/account/stateplan17/nmstateplanseclt r.html

U.S. Department of Education. (2017d). *New Mexico preliminary determination letter*. Retrieved from https://www2.ed.gov/admins/lead/account/stateplan17/nmprelimdetermltr.pdf

U.S. Department of Education. (2017e). *State plan peer review criteria*. Retrieved from https://www2.ed.gov/admins/lead/account/stateplan17/essastateplanpeerreviewcriteria.pdf

Van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M. J., & Holland, R. W. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal, 47*(2), 497-527. doi: 10.3102/0002831209353594

van der Scheer, E. A., Bijlsma, H. J. E., & Glas, C. A. W. (2019). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice, 30*(1), 30-50. doi: 10.1080/09243453.2018.1539015

Vergara (Plaintiff) v. California (Defendent), No. BC484642. (2012). Superior Court of California, County of Los Angeles.

Villegas, A. M. (1988). School failure and cultural mismatch: Another view. *Urban Review, 20*(4), 253-265. doi: 10.1007/bf01120137

Wagner, W., Gollner, R., Helmke, A., Trautwein, U., & Ludtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction 28*(x), 1-11. doi: 10.1016/j.learninstruc.2013.03.003

Wainer, H. (2004). Introduction to a special issue of the *Journal of Educational and Behavioral Statistics* on value-added assessment. *Journal of Educational and Behavioral Statistics, 29*(1), 1-3. doi: 10.3102/10769986029001001

Walker, T. (2015, December 9). With passage of Every Student Succeeds Act, life after NCLB begins. *NEA Today*. Retrieved from http://neatoday.org/2015/12/09/every-student-succeeds-act/

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod student perception survey. *American Educational Research Journal, 53*(6), 1834-1868. doi: 10.3102/0002831216671864

Walsh, K., Joseph, N., Lakis, K., & Lubell, S. (2017). *Running in place: How new teacher evaluations fail to live up to promises*. Washington, DC: National Council on Teacher Quality.

Waxman, H. C., & Eash, M. J. (1983). Utilizing students' perceptions and context variables to analyze effective teaching: A process-product investigation. *Journal of Educational Research, 76*(6), 321-325. doi: 10.1080/00220671.1983.10885476

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student academic achievement: A review. *Review of Educational Research, 73*(1), 89-122. doi: 10.3102/00346543073001089

Weber, N. D., Waxman, H. C., Brown, D. B., & Kelly, L. J. (2016). Informing teacher education through the use of multiple classroom observation instruments. *Teacher Education Quarterly, 43*(1), 91-106.

Weisberg, D., Sexton, S., Mulhearn, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project.

Weiss, E. (2014). Mismatches in Race to the Top limit education improvement. *Education Digest, 79*(5), 60-65.

Weiss, S. (2003). *Highlights from the 2003 National Forum on Education Policy: Nation at Risk continues to affect education system*. Denver, CO: Education Commission of the States.

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin, 127*(6), 797-826. doi: 10.1037/0033-2909.127.6.797

Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35*(3), 521.551. doi: 10.1111/j.1744-6570.1982.tb02208.x

White, M., & Rowan, B. (2014). *User guide to the Measures of Effective Teaching Longitudinal Database (MET LDB).* Ann Arbor, MI: Inter-University Consortium for Political and Social Research.

Whitehurst, G. J. R., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations: Lessons learned in four districts*. Washington, DC: Brown Center on Education Policy at Brookings.

Wiggins, G. (2011). Giving students a voice: The power of feedback to improve teaching. *Educational Horizons, 89*(3), 23-26.

Wilkerson, D. J., Manatt, R. P., Rogers, M. A., & Maughan, R. (2000). Validation of student, principal and self-ratings in 360° feedback® for teacher evaluation. *Journal of Personnel Evaluation in Education, 14*(2), 179-192.

Wise, A. E., Darling-Hammond, L., McLaughlin, M. W., & Bernstein, H. T. (1984). *Teacher evaluation: A study of effective practices*. Santa Monica, CA: RAND Corporation.

Wodtke, G. T. (2012). The impact of education on intergroup attitudes: A multiracial analysis. *Social Psychology Quarterly, 75*(1)*,* 80-106. doi: 10.1177/0190272511430234

Wong, K. K., & Nicotera, A. C. (2004). "Brown v. Board of Education" and the Coleman Report: Social science research and the debate on educational equality. *Peabody Journal of Education, 79*(2), 122-135. doi: 10.1207/s15327930pje7902_8

Worrell, F. C., & Kuterbach, L. D. (2001). The use of student ratings of teacher behaviors with academically talented high school students. *Journal of Secondary Gifted Education, 14*(4), 236-247.

Wright, A. C. (2016, March). *Teachers' perceptions of students' disruptive behavior: The effect of racial congruence and consequences for school suspension*. Paper presented at the Annual Conference of the Association for Education Finance and Policy (AEFP), Denver, CO.

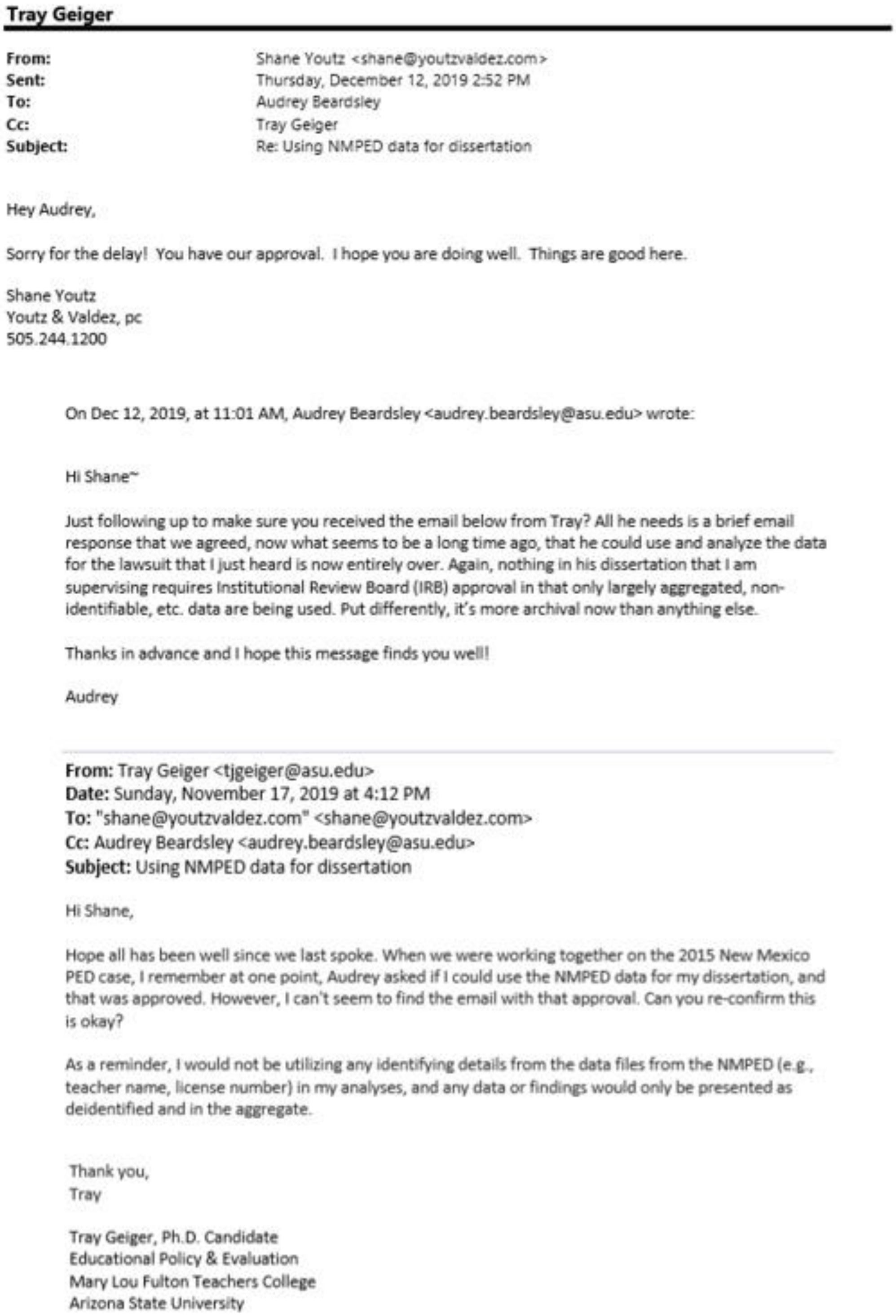
Wright, S. P., White, J. T., Sanders, W. L., & Rivers, J. C. (2010). *SAS EVAAS statistical models*. Cary, NC: SAS Institute, Inc.

Yeh, S. S. (2013). A re-analysis of the effects of teacher replacement using value-added modeling. *Teachers College Record, 115*(1)

Youngs, P., & Grissom, J. A. (2016). Multiple measures in teacher evaluation: Lessons learned and guidelines for practice. In J. A. Grissom & P. Youngs (Eds.), *Improving teacher evaluation systems: Making the most of multiple measures* (pp. 169-183). New York, NY: Teachers College Press.

APPENDIX A

PERMISSION FOR DATA USE



Tray Geiger

From:	Shane Youtz <shane@youtzvaldez.com>
Sent:	Thursday, December 12, 2019 2:52 PM
To:	Audrey Beardsley
Cc:	Tray Geiger
Subject:	Re: Using NMPED data for dissertation

Hey Audrey,

Sorry for the delay! You have our approval. I hope you are doing well. Things are good here.

Shane Youtz
Youtz & Valdez, pc
505.244.1200

On Dec 12, 2019, at 11:01 AM, Audrey Beardsley <audrey.beardsley@asu.edu> wrote:

Hi Shane~

Just following up to make sure you received the email below from Tray? All he needs is a brief email response that we agreed, now what seems to be a long time ago, that he could use and analyze the data for the lawsuit that I just heard is now entirely over. Again, nothing in his dissertation that I am supervising requires Institutional Review Board (IRB) approval in that only largely aggregated, non-identifiable, etc. data are being used. Put differently, it's more archival now than anything else.

Thanks in advance and I hope this message finds you well!

Audrey

From: Tray Geiger <tjgeiger@asu.edu>
Date: Sunday, November 17, 2019 at 4:12 PM
To: "shane@youtzvaldez.com" <shane@youtzvaldez.com>
Cc: Audrey Beardsley <audrey.beardsley@asu.edu>
Subject: Using NMPED data for dissertation

Hi Shane,

Hope all has been well since we last spoke. When we were working together on the 2015 New Mexico PED case, I remember at one point, Audrey asked if I could use the NMPED data for my dissertation, and that was approved. However, I can't seem to find the email with that approval. Can you re-confirm this is okay?

As a reminder, I would not be utilizing any identifying details from the data files from the NMPED (e.g., teacher name, license number) in my analyses, and any data or findings would only be presented as deidentified and in the aggregate.

Thank you,
Tray

Tray Geiger, Ph.D. Candidate
Educational Policy & Evaluation
Mary Lou Fulton Teachers College
Arizona State University

APPENDIX B

REGRESSION MODELS AND LIST OF COVARIATES

Table B1

Regression Model Components for VAS Models

Model Number	Criterion Variable	Main Predictor Variable	Model Covariates
1a	VAS scores	Percent of SE students at a school	Teacher grade level Teacher years of experience Pct. of gifted students – Teacher's classes Pct. of SE students – Teacher's classes Pct. of FRL students – Teacher's classes Pct. of URM students – Teacher's classes Pct. of ELL students – Teacher's school Pct. of FRL students – Teacher's school Pct. of URM students – Teacher's school
1b	VAS scores	Percent of ELL students at a school	Teacher grade level Teacher years of experience Teacher ethnicity Pct. of gifted students – Teacher's classes Pct. of SE students – Teacher's classes Pct. of FRL students – Teacher's classes Pct. of URM students – Teacher's classes Pct. of SE students – Teacher's school Pct. of FRL students – Teacher's school Pct. of URM students – Teacher's school
1c	VAS scores	Percent of FRL students at a school	Teacher years of experience Pct. of gifted students – Teacher's classes Pct. of SE students – Teacher's classes Pct. of FRL students – Teacher's classes Pct. of URM students – Teacher's classes Student enrollment – Teacher's school Pct. of SE students – Teacher's school Pct. of ELL students – Teacher's school Pct. of URM students – Teacher's school
1d	VAS scores	Percent of URM students at a school	Teacher grade level Teacher ethnicity Teacher years of experience Pct. of gifted students – Teacher's classes Pct. of SE students – Teacher's classes Pct. of FRL students – Teacher's classes Pct. of URM students – Teacher's classes Student enrollment – Teacher's school Pct. of SE students – Teacher's school Pct. of ELL students – Teacher's school Pct. of FRL students – Teacher's school

Table B2

Regression Model Components for Models with Classroom Observation Scores as the Main Outcome Variable

Model Number	Dependent Variable	Main Independent Variable	Model Controls
2a	Observation scores	Percent of SE students at a school	Teacher grade level Teacher years of experience Pct. of gifted students – Teacher's classes Pct. of SE students – Teacher's classes Pct. of FRL students – Teacher's classes Pct. of URM students – Teacher's classes Pct. of ELL students – Teacher's school Pct. of FRL students – Teacher's school Pct. of URM students – Teacher's school
2b	Observation scores	Percent of ELL students at a school	Teacher grade level Teacher ethnicity Teacher years of experience Pct. of gifted students – Teacher's classes Pct. of SE students – Teacher's classes Pct. of FRL students – Teacher's classes Pct. of URM students – Teacher's classes Pct. of SE students – Teacher's school Pct. of FRL students – Teacher's school Pct. of URM students – Teacher's school
2c	Observation scores	Percent of FRL students at a school	Teacher years of experience Pct. of gifted students – Teacher's classes Pct. of SE students – Teacher's classes Pct. of FRL students – Teacher's classes Pct. of URM students – Teacher's classes Student enrollment – Teacher's school Pct. of SE students – Teacher's school Pct. of ELL students – Teacher's school Pct. of URM students – Teacher's school
2d	Observation scores	Percent of URM students at a school	Teacher grade level Teacher ethnicity Teacher years of experience Pct. of gifted students – Teacher's classes Pct. of SE students – Teacher's classes Pct. of FRL students – Teacher's classes Pct. of URM students – Teacher's classes Student enrollment – Teacher's school Pct. of SE students – Teacher's school Pct. of ELL students – Teacher's school Pct. of FRL students – Teacher's school

Table B3

Regression Model Components for Models with PPP Scores as the Main Outcome Variable

Model Number	Dependent Variable	Main Independent Variable	Model Controls
3a	PPP scores	Percent of SE students at a school	Teacher grade level Teacher years of experience Pct. of gifted students – Teacher's classes Pct. of SE students – Teacher's classes Pct. of FRL students – Teacher's classes Pct. of URM students – Teacher's classes Pct. of ELL students – Teacher's school Pct. of FRL students – Teacher's school Pct. of URM students – Teacher's school
3b	PPP scores	Percent of ELL students at a school	Teacher grade level Teacher ethnicity Teacher years of experience Pct. of gifted students – Teacher's classes Pct. of SE students – Teacher's classes Pct. of FRL students – Teacher's classes Pct. of URM students – Teacher's classes Pct. of SE students – Teacher's school Pct. of FRL students – Teacher's school Pct. of URM students – Teacher's school
3c	PPP scores	Percent of FRL students at a school	Teacher years of experience Pct. of gifted students – Teacher's classes Pct. of SE students – Teacher's classes Pct. of FRL students – Teacher's classes Pct. of URM students – Teacher's classes Student enrollment – Teacher's school Pct. of SE students – Teacher's school Pct. of ELL students – Teacher's school Pct. of URM students – Teacher's school
3d	PPP scores	Percent of URM students at a school	Teacher grade level Teacher ethnicity Teacher years of experience Pct. of gifted students – Teacher's classes Pct. of SE students – Teacher's classes Pct. of FRL students – Teacher's classes Pct. of URM students – Teacher's classes Student enrollment – Teacher's school Pct. of SE students – Teacher's school Pct. of ELL students – Teacher's school Pct. of FRL students – Teacher's school

Table B4

Regression Model Components for Models with SPS Scores as the Main Outcome Variable

Model Number	Dependent Variable	Main Independent Variable	Model Controls
4a	SPS scores	Percent of SE students at a school	Teacher grade level Teacher years of experience Pct. of gifted students – Teacher's classes Pct. of SE students – Teacher's classes Pct. of FRL students – Teacher's classes Pct. of URM students – Teacher's classes Pct. of ELL students – Teacher's school Pct. of FRL students – Teacher's school Pct. of URM students – Teacher's school
4b	SPS scores	Percent of ELL students at a school	Teacher grade level Teacher ethnicity Teacher years of experience Pct. of gifted students – Teacher's classes Pct. of SE students – Teacher's classes Pct. of FRL students – Teacher's classes Pct. of URM students – Teacher's classes Pct. of SE students – Teacher's school Pct. of FRL students – Teacher's school Pct. of URM students – Teacher's school
4c	SPS scores	Percent of FRL students at a school	Teacher years of experience Pct. of gifted students – Teacher's classes Pct. of SE students – Teacher's classes Pct. of FRL students – Teacher's classes Pct. of URM students – Teacher's classes Student enrollment – Teacher's school Pct. of SE students – Teacher's school Pct. of ELL students – Teacher's school Pct. of URM students – Teacher's school
4d	SPS scores	Percent of URM students at a school	Teacher grade level Teacher ethnicity Teacher years of experience Pct. of gifted students – Teacher's classes Pct. of SE students – Teacher's classes Pct. of FRL students – Teacher's classes Pct. of URM students – Teacher's classes Student enrollment – Teacher's school Pct. of SE students – Teacher's school Pct. of ELL students – Teacher's school Pct. of FRL students – Teacher's school

Table B5

List of Teacher-, Classroom-, and School-Level Control Variables

Teacher-Level Variables	Classroom-Level Variables	School-Level Variables
Grade Level Taught	Pct. of Gifted Students	Pct. of SE Students
Years of Experience	Pct. of SE Students	Pct. of ELL Students
Race/Ethnicity	Pct. of FRL Students	Pct. of FRL Students
	Pct. of URM Students	Pct. of URM Students
		Student Enrollment

Note: The percent of classroom-level and school-level URM students was derived by summing the total number of URM students, per classroom and per school, respectively, and dividing that by the total number of all students, per classroom and per school, respectively and then multiplying by 100.

APPENDIX C

REGRESSION MODEL OUTPUTS

Table C1

Model 1a Output, All Years (VAS as Criterion Variable)

	Year 1		Year 2		Year 3	
	2013-2014		2014-2015		2015-2016	
n	2,733		2,738		8,963	
F	8.97		8.17		32.23	
df	13, 2,719		11, 2,726		11, 8,951	
p	0.000		0.000		0.000	
R^2	0.0317		0.0273		0.0378	
	Coef. ^	*SE* ^	Coef. ^	*SE* ^	Coef. ^	*SE* ^
Pct. of SE Students – Teacher's School (Per 10%)	__3.3**__	***1.00***	__3.4**__	***0.99***	__1.2*__	***0.59***
Teacher School Level Taught (Elem. as base)						
Middle	__-3.1**__	***1.02***	1.6	*1.13*	*N/A – Not in model*	
High	__3.0*__	***1.32***	1.8	*1.28*		
Teacher Years of Experience	0.0	*0.46*	0.0	*0.05*	__0.2***__	***0.03***
Pct. of Gifted Students – Teacher's Classes (Per 10%)	__1.9***__	***0.44***	1.0	*0.56*	__1.2***__	***0.25***
Pct. of SE Students – Teacher's Classes (Per 10%)	0.2	*0.20*	0.3	*0.22*	0.1	*0.11*
Pct. of FRL Students – Teacher's Classes (Per 10%)	0.0	*0.53*	-0.9	*0.64*	-0.4	*0.20*
Pct. of URM Students – Teacher's Classes (Per 10%)	-1.0	*0.57*	0.1	*0.57*	-0.3	*0.31*
Pct. of ELL Students – Teacher's School (Per 10%)	__1.2**__	***0.41***	__2.3***__	***0.45***	0.5	*0.26*
Pct. of FRL Students – Teacher's School (Per 10%)	-0.7	*0.59*	0.6	*0.66*	__-0.6**__	***0.22***
Pct. of URM Students – Teacher's School (Per 10%)	0.4	*0.72*	__-2.0**__	***0.68***	__-0.8*__	***0.37***
Constant	__52.3***__	***2.64***	__59.2***__	***2.70***	__62.9***__	***1.46***

*** $p < 0.001$ ** $p \leq 0.01$ * $p \leq 0.05$

^ All units are expressed in percentages.

Note: The main predictor variable of interest (i.e., Percent of SE students within a teacher's school) is highlighted in light gray.

Note: Any coefficients that appear to equal zero percent are actually greater than or less than zero percent, but due to rounding rules appear to equal zero percent.

Note: Teacher School Level Taught is omitted from the Year 3 (2015-2016) model as it was not present in the data provided by the NMPED.

Table C2

Model 1b Output, All Years (VAS as Criterion Variable)

	Year 1 2013-2014		Year 2 2014-2015		Year 3 2015-2016	
n	2,733		2,738		8,963	
F	8.92		8.35		31.68	
df	14, 2,718		12, 2,725		12, 8,950	
p	0.000		0.000		0.000	
R^2	0.0334		0.0298		0.0389	
	Coef. ^	*SE* ^	Coef. ^	*SE* ^	Coef. ^	*SE* ^
Pct. of ELL Students – Teacher's School (Per 10%)	**1.1****	***0.41***	**2.2*****	***0.45***	0.5	*0.26*
Teacher School Level Taught (Elem. as base)						
Middle	**-3.1****	***1.02***	1.8	*1.13*	*N/A – Not in model*	
High	**3.0***	***1.32***	2.0	*1.28*		
Teacher URM (No as base)	**1.9***	***0.91***	**2.5****	***0.94***	**1.8****	***0.55***
Teacher Years of Experience	0.0	*0.05*	0.0	*0.05*	**0.2*****	***0.03***
Pct. of Gifted Students – Teacher's Classes (Per 10%)	**1.9*****	***0.43***	1.0	*0.56*	**1.3*****	***0.25***
Pct. of SE Students – Teacher's Classes (Per 10%)	0.2	*0.20*	0.3	*0.22*	0.1	*0.11*
Pct. of FRL Students – Teacher's Classes (Per 10%)	0.0	*0.53*	-0.9	*0.64*	**-0.4***	***0.20***
Pct. of URM Students – Teacher's Classes (Per 10%)	-1.0	*0.57*	0.0	*0.57*	-0.3	*0.31*
Pct. of SE Students – Teacher's School (Per 10%)	**3.2****	***1.01***	**3.2****	***0.98***	**1.2***	***0.59***
Pct. of FRL Students – Teacher's School (Per 10%)	-0.6	*0.59*	0.5	*0.66*	**-0.6****	***0.22***
Pct. of URM Students – Teacher's School (Per 10%)	0.4	*0.72*	**-1.9****	***0.68***	**-0.8***	***0.37***
Constant	**51.9*****	***2.64***	**58.7*****	***2.71***	**63.0*****	***1.46***

*** $p < 0.001$ ** $p \leq 0.01$ * $p \leq 0.05$

^ All units are expressed in percentages.

Note: The main predictor variable of interest (i.e., Percent of ELL students within a teacher's school) is highlighted in light gray.

Note: Any coefficients that appear to equal zero percent are actually greater than or less than zero percent, but due to rounding rules appear to equal zero percent.

Note: Teacher School Level Taught is omitted from the Year 3 (2015-2016) model as it was not present in the data provided by the NMPED.

Table C3

Model 1c Output, All Years (VAS as Criterion Variable)

	Year 1 2013-2014		Year 2 2014-2015		Year 3 2015-2016	
n	2,733		2,738		8,963	
F	7.75		9.11		32.32	
df	12, 2,720		10, 2,727		12, 8,950	
p	0.000		0.000		0.000	
R^2	0.0256		0.0280		0.0399	
	Coef. ^	*SE* ^	Coef. ^	*SE* ^	Coef. ^	*SE* ^
Pct. of FRL Students – Teacher's School (Per 10%)	-0.6	*0.59*	0.6	*0.66*	**-0.8*** **	***0.23***
Teacher Years of Experience	0.0	*0.05*	0.0	*0.05*	**0.2*** **	***0.03***
Pct. of Gifted Students – Teacher's Classes (Per 10%)	**1.8*** **	***0.43***	1.1	*0.56*	**1.2*** **	***0.25***
Pct. of SE Students – Teacher's Classes (Per 10%)	0.1	*0.20*	0.3	*0.22*	0.1	*0.11*
Pct. of FRL Students – Teacher's Classes (Per 10%)	0.1	*0.52*	-0.9	*0.64*	**-0.5* **	***0.21***
Pct. of URM Students – Teacher's Classes (Per 10%)	-1.1	*0.56*	0.1	*0.57*	-0.2	*0.31*
Student Enrollment – Teacher's School (Per 100 Students)	**0.3** **	***0.11***	**0.3* **	***0.12***	**-0.3*** **	***0.06***
Pct. of SE Students – Teacher's School (Per 10%)	**3.5*** **	***1.00***	**3.6*** **	***0.98***	1.0	*0.59*
Pct. of ELL Students – Teacher's School (Per 10%)	**1.1** **	***0.39***	**2.2*** **	***0.42***	0.4	*0.27*
Pct. of URM Students – Teacher's School (Per 10%)	0.6	*0.71*	**-2.0** **	***0.67***	-0.5	*0.38*
Constant	**49.3*** **	***2.89***	**57.9*** **	***2.81***	**64.9*** **	***1.52***

*** $p < 0.001$ ** $p \leq 0.01$ * $p \leq 0.05$

^ All units are expressed in percentages.

Note: The main predictor variable of interest (i.e., Percent of FRL students within a teacher's school) is highlighted in light gray.

Note: Any coefficients that appear to equal zero percent are actually greater than or less than zero percent, but due to rounding rules appear to equal zero percent.

Table C4

Model 1d Output, All Years (VAS as Criterion Variable)

	Year 1 2013-2014		Year 2 2014-2015		Year 3 2015-2016	
n	2,733		2,738		8,963	
F	8.32		7.78		30.94	
df	15, 2,717		13, 2,724		13, 8,949	
p	0.000		0.000		0.000	
R^2	0.0339		0.0307		0.0409	
	Coef. ^	*SE* ^	Coef. ^	*SE* ^	Coef. ^	*SE* ^
Pct. of URM Students – Teacher's School (Per 10%)	0.4	0.27	**-1.9****	**0.68**	0.5	0.38
Teacher School Level Taught (Elem. as base)						
Middle	**-3.2****	**1.02**	1.6	1.14	*N/A – Not in*	
High	2.2	1.57	0.8	1.52	*model*	
Teacher URM (No as base)	**2.0***	**0.91**	**2.4***	**0.94**	**1.7****	**0.55**
Teacher Years of Experience	0.0	0.05	0.0	0.05	**0.2*****	**0.03**
Pct. of Gifted Students – Teacher's Classes (Per 10%)	**1.9*****	**0.44**	1.0	0.56	**1.2*****	**0.25**
Pct. of SE Students – Teacher's Classes (Per 10%)	0.2	0.20	0.3	0.22	0.1	0.11
Pct. of FRL Students – Teacher's Classes (Per 10%)	0.0	0.53	-0.9	0.64	**-0.5****	**0.21**
Pct. of URM Students – Teacher's Classes (Per 10%)	-1.0	0.57	0.0	0.57	0.2	0.31
Student Enrollment – Teacher's School (Per 100 Students)	0.1	0.14	0.2	0.14	**-0.3*****	**0.06**
Pct. of SE Students – Teacher's School (Per 10%)	**3.3****	**1.02**	**3.4****	**0.99**	1.0	0.59
Pct. of ELL Students – Teacher's School (Per 10%)	**1.0***	**0.43**	**2.1*****	**0.46**	0.3	0.27
Pct. of FRL Students – Teacher's School (Per 10%)	-0.5	0.59	0.6	0.66	**-0.8*****	**0.23**
Constant	**50.6*****	**2.88**	**57.3*****	**2.81**	**65.0*****	**1.52**

*** $p < 0.001$ ** $p \leq 0.01$ * $p \leq 0.05$

^ All units are expressed in percentages.

Note: The main predictor variable of interest (i.e., Percent of URM students within a teacher's school) is highlighted in light gray.

Note: Any coefficients that appear to equal zero percent are actually greater than or less than zero percent, but due to rounding rules appear to equal zero percent.

Note: Teacher School Level Taught is omitted from the Year 3 (2015-2016) model as it was not present in the data provided by the NMPED.

Table C5

Model 2a Output, All Years (Observations as Criterion Variable)

	Year 1 2013-2014		Year 2 2014-2015		Year 3 2015-2016	
n	2,733		2,738		8,963	
F	20.60		23.68		56.20	
df	13, 2,719		11, 2,726		11, 8,951	
p	0.000		0.000		0.000	
R^2	0.0825		0.0844		0.0618	
	Coef. ^	*SE* ^	Coef. ^	*SE* ^	Coef. ^	*SE* ^
Pct. of SE Students – Teacher's School (Per 10%)	-0.4	0.39	-0.5	0.42	0.4	0.23
Teacher School Level Taught (Elem. as base)						
Middle	**-1.5*****	**0.41**	**-2.4*****	**0.47**	*N/A – Not in model*	
High	**-2.7*****	**0.49**	**-3.4*****	**0.51**		
Teacher Years of Experience	**0.1*****	**0.02**	**0.1*****	**0.02**	**0.1*****	**0.01**
Pct. of Gifted Students – Teacher's Classes (Per 10%)	0.3	0.17	**0.6***	**0.23**	**0.4*****	**0.10**
Pct. of SE Students – Teacher's Classes (Per 10%)	-0.1	0.07	0.0	0.11	**-0.2*****	**0.05**
Pct. of FRL Students – Teacher's Classes (Per 10%)	**-0.7****	**0.21**	**-0.6****	**0.27**	**-0.4*****	**0.08**
Pct. of URM Students – Teacher's Classes (Per 10%)	0.0	0.23	-0.3	0.23	**-0.4*****	**0.13**
Pct. of ELL Students – Teacher's School (Per 10%)	**-0.5****	**0.15**	-0.3	0.17	0.1	0.10
Pct. of FRL Students – Teacher's School (Per 10%)	0.1	0.22	0.3	0.28	**0.2***	**0.09**
Pct. of URM Students – Teacher's School (Per 10%)	0.0	0.28	-0.2	0.27	-0.1	0.15
Constant	**73.5*****	**1.00**	**77.9*****	**1.12**	**75.9*****	**0.56**

*** $p < 0.001$ ** $p \leq 0.01$ * $p \leq 0.05$

^ All units are expressed in percentages.

Note: The main predictor variable of interest (i.e., Percent of SE students within a teacher's school) is highlighted in light gray.

Note: Any coefficients that appear to equal zero percent are actually greater than or less than zero percent, but due to rounding rules appear to equal zero percent.

Note: Teacher School Level Taught is omitted from the Year 3 (2015-2016) model as it was not present in the data provided by the NMPED.

Table C6

Model 2b Output, All Years (Observations as Criterion Variable)

	Year 1 2013-2014		Year 2 2014-2015		Year 3 2015-2016	
n	2,733		2,738		8,963	
F	19.31		22.61		52.35	
df	14, 2,718		12, 2,725		12, 8,950	
p	0.000		0.000		0.000	
R^2	0.0834		0.0873		0.0625	
	Coef. ^	*SE* ^	Coef. ^	*SE* ^	Coef. ^	*SE* ^
Pct. of ELL Students – Teacher's School (Per 10%)	**-0.5****	**0.15**	-0.3	0.17	0.0	0.10
Teacher School Level Taught (Elem. as base)						
Middle	**-1.5*****	**0.41**	**-2.3*****	**0.47**	*N/A – Not in*	
High	**-2.8*****	**0.49**	**-3.3*****	**0.51**	*model*	
Teacher URM (No as base)	-0.6	0.35	**1.1****	**0.38**	**0.6****	**0.22**
Teacher Years of Experience	**0.1*****	**0.02**	**0.1*****	**0.02**	**0.1*****	**0.01**
Pct. of Gifted Students – Teacher's Classes (Per 10%)	0.3	0.18	**0.6***	**0.23**	**0.4*****	**0.10**
Pct. of SE Students – Teacher's Classes (Per 10%)	-0.1	0.07	0.0	0.11	**-0.2*****	0.05
Pct. of FRL Students – Teacher's Classes (Per 10%)	**-0.7****	**0.21**	**-0.7****	**0.27**	**-0.4*****	**0.08**
Pct. of URM Students – Teacher's Classes (Per 10%)	0.0	0.23	-0.4	0.23	**-0.5*****	**0.13**
Pct. of SE Students – Teacher's School (Per 10%)	-0.4	0.39	-0.6	0.41	0.4	0.23
Pct. of FRL Students – Teacher's School (Per 10%)	0.1	0.22	0.3	0.28	**0.2***	**0.09**
Pct. of URM Students – Teacher's School (Per 10%)	0.0	0.28	-0.1	0.27	-0.1	0.15
Constant	**73.7*****	**1.00**	**77.7*****	**1.10**	**75.9*****	**0.56**

*** $p < 0.001$ ** $p \leq 0.01$ * $p \leq 0.05$

^ All units are expressed in percentages.

Note: The main predictor variable of interest (i.e., Percent of ELL students within a teacher's school) is highlighted in light gray.

Note: Any coefficients that appear to equal zero percent are actually greater than or less than zero percent, but due to rounding rules appear to equal zero percent.

Note: Teacher School Level Taught is omitted from the Year 3 (2015-2016) model as it was not present in the data provided by the NMPED.

Table C7

Model 2c Output, All Years (Observations as Criterion Variable)

	Year 1 2013-2014		Year 2 2014-2015		Year 3 2015-2016	
n	2,733		2,738		8,963	
F	19.58		21.65		52.31	
df	12, 2,720		10, 2,727		12, 8,950	
p	0.000		0.000		0.000	
R^2	0.0722		0.0729		0.0619	
	Coef. ^	*SE* ^	Coef. ^	*SE* ^	Coef. ^	*SE* ^
Pct. of FRL Students – Teacher's School (Per 10%)	0.2	0.22	0.3	0.28	0.2	0.09
Teacher Years of Experience	0.1**	0.02	0.1***	0.02	0.1***	0.01
Pct. of Gifted Students – Teacher's Classes (Per 10%)	0.4	0.18	0.5*	0.23	0.4*** -0.2***	0.10
Pct. of SE Students – Teacher's Classes (Per 10%)	-0.1	0.07	0.1	0.11		0.05
Pct. of FRL Students – Teacher's Classes (Per 10%)	-0.6**	0.21	-0.7*	0.27	-0.5***	0.08
Pct. of URM Students – Teacher's Classes (Per 10%)	-0.1	0.23	-0.4	0.23	-0.4***	0.13
Student Enrollment – Teacher's School (Per 100 Students)	0.1	0.04	0.0***	0.00	0.0	0.02
Pct. of SE Students – Teacher's School (Per 10%)	0.4	0.39	-0.2	0.43	0.4	0.23
Pct. of ELL Students – Teacher's School (Per 10%)	-0.0%	0.15	0.1	0.16	0.1	0.10
Pct. of URM Students – Teacher's School (Per 10%)	-0.2	0.28	-0.3	0.27	-0.1	0.15
Constant	71.3***	1.06	77.9***	1.19	76.1***	0.58

*** $p < 0.001$ ** $p \leq 0.01$ * $p \leq 0.05$

^ All units are expressed in percentages.

Note: The main predictor variable of interest (i.e., Percent of FRL students within a teacher's school) is highlighted in light gray.

Note: Any coefficients that appear to equal zero percent are actually greater than or less than zero percent, but due to rounding rules appear to equal zero percent.

Table C8

Model 2d Output, All Years (Observations as Criterion Variable)

	Year 1 2013-2014		Year 2 2014-2015		Year 3 2015-2016	
n	2,733		2,738		8,963	
F	18.93		21.19		49.03	
df	15, 2,717		13, 2,724		13, 8,949	
p	0.000		0.000		0.000	
R^2	0.0917		0.0882		0.0626	
	Coef. ^	*SE* ^	Coef. ^	*SE* ^	Coef. ^	*SE* ^
Pct. of URM Students – Teacher's School (Per 10%)	0.0	0.27	-0.1	0.27	-0.1	0.15
Teacher School Level Taught (Elem. as base)					*N/A – Not in model*	
Middle	**-1.7*****	**0.42**	**-2.3*****	**0.47**		
High	**-4.2*****	**0.58**	**-2.8*****	**0.62**		
Teacher URM (No as base)	-0.6	0.35	**1.2****	**0.38**	**0.6****	**0.22**
Teacher Years of Experience	**0.1*****	**0.02**	**0.1*****	**0.02**	**0.1*****	**0.01**
Pct. of Gifted Students – Teacher's Classes (Per 10%)	0.3	0.18	**0.6***	**0.23**	**0.4*****	**0.10**
Pct. of SE Students – Teacher's Classes (Per 10%)	-0.1	0.07	0.0	0.11	**-0.2*****	**0.05**
Pct. of FRL Students – Teacher's Classes (Per 10%)	**-0.7****	**0.21**	**-0.7****	**0.27**	**-0.5*****	**0.08**
Pct. of URM Students – Teacher's Classes (Per 10%)	0.1	0.23	-0.4	0.23	**-0.5*****	**0.13**
Student Enrollment – Teacher's School (Per 100 Students)	**0.2*****	**0.05**	0.0	0.00	0.0	0.02
Pct. of SE Students – Teacher's School (Per 10%)	-0.1	0.39	-0.6	0.41	0.4	0.23
Pct. of ELL Students – Teacher's School (Per 10%)	**-0.6*****	**0.16**	-0.3	0.18	0.0	0.10
Pct. of FRL Students – Teacher's School (Per 10%)	0.2	0.22	0.2	0.27	0.2	0.09
Constant	**71.6*****	**1.05**	**78.2*****	**1.15**	**76.1*****	**0.58**

*** $p < 0.001$ ** $p \leq 0.01$ * $p \leq 0.05$

^ All units are expressed in percentages.

Note: The main predictor variable of interest (i.e., Percent of URM students within a teacher's school) is highlighted in light gray.

Note: Any coefficients that appear to equal zero percent are actually greater than or less than zero percent, but due to rounding rules appear to equal zero percent.

Note: Teacher School Level Taught is omitted from the Year 3 (2015-2016) model as it was not present in the data provided by the NMPED.

Table C9

Model 3a Output, All Years (PPP as Criterion Variable)

	Year 1		Year 2		Year 3	
	2013-2014		2014-2015		2015-2016	
n	2,733		2,738		8,963	
F	21.88		28.71		48.70	
df	13, 2,719		11, 2,726		11, 8,951	
p	0.000		0.000		0.000	
R^2	0.0932		0.1020		0.0546	
	Coef. ^	*SE* ^	Coef. ^	*SE* ^	Coef. ^	*SE* ^
Pct. of SE Students – Teacher's School (Per 10%)	0.5	0.46	-0.4	0.45	**1.0*****	**0.26**
Teacher School Level Taught (Elem. as base)						
Middle	**-1.3****	**0.47**	**-2.3*****	**0.51**	*N/A – Not in model*	
High	**-1.6****	**0.56**	**-4.6*****	**0.56**		
Teacher Years of Experience	**0.1*****	**0.02**	**0.1*****	**0.02**	**0.1*****	**0.01**
Pct. of Gifted Students – Teacher's Classes (Per 10%)	0.2	0.20	0.1	0.26	**0.3****	**0.12**
Pct. of SE Students – Teacher's Classes (Per 10%)	-0.1	0.09	0.1	0.13	**-0.2****	**0.06**
Pct. of FRL Students – Teacher's Classes (Per 10%)	**-0.8****	**0.24**	**-0.7***	**0.28**	**-0.4*****	**0.09**
Pct. of URM Students – Teacher's Classes (Per 10%)	-0.1	0.27	-0.4	0.25	**-0.5****	**0.14**
Pct. of ELL Students – Teacher's School (Per 10%)	-0.3	0.18	-0.3	0.19	0.2	0.12
Pct. of FRL Students – Teacher's School (Per 10%)	0.0	0.26	0.2	0.28	0.2	0.10
Pct. of URM Students – Teacher's School (Per 10%)	-0.1	0.32	-0.3	0.30	**-0.4***	**0.17**
Constant	**76.4*****	**1.18**	**83.9*****	**1.24**	**78.4*****	**0.65**

*** $p < 0.001$ ** $p \leq 0.01$ * $p \leq 0.05$

^ All units are expressed in percentages.

Note: The main predictor variable of interest (i.e., Percent of SE students within a teacher's school) is highlighted in light gray.

Note: Any coefficients that appear to equal zero percent are actually greater than or less than zero percent, but due to rounding rules appear to equal zero percent.

Note: Teacher School Level Taught is omitted from the Year 3 (2015-2016) model as it was not present in the data provided by the NMPED.

Table C10

Model 3b Output, All Years (PPP as Criterion Variable)

	Year 1 2013-2014		Year 2 2014-2015		Year 3 2015-2016	
n	2,733		2,738		8,963	
F	20.54		26.81		44.72	
df	14, 2,718		12, 2,725		12, 8,950	
p	0.000		0.000		0.000	
R^2	0.0939		0.1036		0.0547	
	Coef. ^	*SE* ^	Coef. ^	*SE* ^	Coef. ^	*SE* ^
Pct. of ELL Students – Teacher's School (Per 10%)	-0.3	0.18	-0.3	0.20	0.2	0.12
Teacher School Level Taught (Elem. as base)						
Middle	-1.3**	0.47	-2.2***	0.51	*N/A – Not in model*	
High	-1.7**	0.56	-4.5***	0.56		
Teacher URM (No as base)	-0.6	0.40	0.9*	0.42	0.1	0.25
Teacher Years of Experience	0.1***	0.02	0.1***	0.02	0.1***	0.01
Pct. of Gifted Students – Teacher's Classes (Per 10%)	0.2	0.20	0.1	0.26	0.3**	0.12
Pct. of SE Students – Teacher's Classes (Per 10%)	-0.0	0.09	-0.1	0.13	-0.2**	0.06
Pct. of FRL Students – Teacher's Classes (Per 10%)	-0.8**	0.24	-0.7**	0.28	-0.4***	0.09
Pct. of URM Students – Teacher's Classes (Per 10%)	-0.1	0.27	-0.5	0.25	-0.5**	0.14
Pct. of SE Students – Teacher's School (Per 10%)	0.5	0.45	-0.5	0.45	1.0***	0.26
Pct. of FRL Students – Teacher's School (Per 10%)	0.0	0.26	0.2	0.28	-0.4	0.10
Pct. of URM Students – Teacher's School (Per 10%)	-0.1	0.32	-0.2	0.30	-0.4*	0.17
Constant	76.5***	1.18	83.8***	1.22	78.4***	0.65

*** $p < 0.001$ ** $p \leq 0.01$ * $p \leq 0.05$

^ All units are expressed in percentages.

Note: The main predictor variable of interest (i.e., Percent of ELL students within a teacher's school) is highlighted in light gray.

Note: Any coefficients that appear to equal zero percent are actually greater than or less than zero percent, but due to rounding rules appear to equal zero percent.

Note: Teacher School Level Taught is omitted from the Year 3 (2015-2016) model as it was not present in the data provided by the NMPED.

Table C11

Model 3c Output, All Years (PPP as Criterion Variable)

	Year 1 2013-2014		Year 2 2014-2015		Year 3 2015-2016	
n	2,733		2,738		8,963	
F	19.58		25.97		45.00	
df	12, 2,720		10, 2,727		12, 8,950	
p	0.000		0.000		0.000	
R^2	0.0946		0.0905		0.0549	
	Coef. ^	*SE* ^	Coef. ^	*SE* ^	Coef. ^	*SE* ^
Pct. of FRL Students – Teacher's School (Per 10%)	0.1	0.26	0.2	0.29	0.2	0.10
Teacher Years of Experience	0.1***	0.02	0.1***	0.02	0.1***	0.01
Pct. of Gifted Students – Teacher's Classes (Per 10%)	0.2	0.20	0.1	0.27	0.3** -0.2**	0.12
Pct. of SE Students – Teacher's Classes (Per 10%)	-0.1	0.09	0.0	0.13		0.06
Pct. of FRL Students – Teacher's Classes (Per 10%)	-0.7**	0.24	-0.7*	0.28	-0.5***	0.10
Pct. of URM Students – Teacher's Classes (Per 10%)	-0.2	0.27	-0.5	0.25	-0.5**	0.14
Student Enrollment – Teacher's School (Per 100 Students)	0.2***	0.04	0.0***	0.00	0.0	0.03
Pct. of SE Students – Teacher's School (Per 10%)	1.3**	0.45	-0.1	0.47	0.9***	0.26
Pct. of ELL Students – Teacher's School (Per 10%)	-0.1	0.17	0.3	0.18	0.2	0.12
Pct. of URM Students – Teacher's School (Per 10%)	-0.3	0.31	-0.5	0.30	-0.4*	0.17
Constant	73.3***	1.26	84.3***	1.35	78.7***	0.69

*** $p < 0.001$ ** $p \leq 0.01$ * $p \leq 0.05$

^ All units are expressed in percentages.

Note: The main predictor variable of interest (i.e., Percent of FRL students within a teacher's school) is highlighted in light gray.

Note: Any coefficients that appear to equal zero percent are actually greater than or less than zero percent, but due to rounding rules appear to equal zero percent.

Table C12

Model 3d Output, All Years (PPP as Criterion Variable)

	Year 1 2013-2014		Year 2 2014-2015		Year 3 2015-2016	
n	2,733		2,738		8,963	
F	21.77		25.02		41.62	
df	15, 2,717		13, 2,725		13, 8,949	
p	0.000		0.000		0.000	
R^2	0.1063		0.1055		0.0549	
	Coef. ^	*SE* ^	Coef. ^	*SE* ^	Coef. ^	*SE* ^
Pct. of URM Students – Teacher's School (Per 10%)	-0.1	0.32	-0.2	0.30	-0.4*	0.17
Teacher School Level Taught (Elem. as base)						
Middle	-1.5**	0.48	-2.0***	0.51	*N/A – Not in model*	
High	-3.6***	0.65	-3.7***	0.66		
Teacher URM (No as base)	-0.6	0.40	1.0*	0.42	0.1	0.25
Teacher Years of Experience	0.1***	0.02	0.1***	0.02	0.1***	0.01
Pct. of Gifted Students – Teacher's Classes (Per 10%)	0.2	0.20	0.1	0.26	0.3**	0.12
Pct. of SE Students – Teacher's Classes (Per 10%)	-0.1	0.09	-0.1	0.13	-0.2**	0.06
Pct. of FRL Students – Teacher's Classes (Per 10%)	-0.8**	0.24	-0.7**	0.28	-0.5***	0.10
Pct. of URM Students – Teacher's Classes (Per 10%)	-0.1	0.26	-0.4	0.25	-0.5**	0.14
Student Enrollment – Teacher's School (Per 100 Students)	0.3***	0.05	0.0*	0.00	0.0	0.03
Pct. of SE Students – Teacher's School (Per 10%)	-0.9	0.45	-0.6	0.44	-0.9***	0.26
Pct. of ELL Students – Teacher's School (Per 10%)	-0.4*	0.18	-0.2	0.20	0.2	0.12
Pct. of FRL Students – Teacher's School (Per 10%)	0.1	0.26	0.2	0.28	0.2	0.10
Constant	73.7***	1.25	84.6***	1.30	78.7***	0.69

*** $p < 0.001$ ** $p \leq 0.01$ * $p \leq 0.05$

^ All units are expressed in percentages.

Note: The main predictor variable of interest (i.e., Percent of URM students within a teacher's school) is highlighted in light gray.

Note: Any coefficients that appear to equal zero percent are actually greater than or less than zero percent, but due to rounding rules appear to equal zero percent.

Note: Teacher School Level Taught is omitted from the Year 3 (2015-2016) model as it was not present in the data provided by the NMPED.

Table C13

Model 4a Output, All Years (SPS as Criterion Variable)

	Year 1 2013-2014		Year 2 2014-2015		Year 3 2015-2016	
n	2,733		2,738		8,963	
F	5.49		52.41		66.65	
df	13, 2,719		11, 2,726		11, 8,951	
p	0.000		0.000		0.000	
R^2	0.0249		.1723		0.0689	
	Coef. ^	*SE* ^	Coef. ^	*SE* ^	Coef. ^	*SE* ^
Pct. of SE Students – Teacher's School (Per 10%)	**-1.9****	**0.61**	**-1.0****	**0.37**	0.1	0.23
Teacher School Level Taught (Elem. as base)						
Middle	**-2.2****	**0.63**	**-5.8*****	**0.42**	*N/A – Not in model*	
High	**-3.9*****	**0.72**	**-8.2*****	**0.50**		
Teacher Years of Experience	**-0.1***	**0.03**	**0.0****	**0.02**	0.0	0.01
Pct. of Gifted Students – Teacher's Classes (Per 10%)	**0.7****	**0.28**	**1.4*****	**0.25**	**-0.8*****	**.011**
Pct. of SE Students – Teacher's Classes (Per 10%)	0.0	0.12	**0.3****	**0.09**	0.1	0.05
Pct. of FRL Students – Teacher's Classes (Per 10%)	0.5	0.30	-0.1	0.25	**0.6*****	**0.09**
Pct. of URM Students – Teacher's Classes (Per 10%)	-0.2	0.33	**-0.7****	**0.22**	**-0.7*****	**0.13**
Pct. of ELL Students – Teacher's School (Per 10%)	-0.3	0.33	**0.8*****	**0.16**	**2.0*****	**0.10**
Pct. of FRL Students – Teacher's School (Per 10%)	0.0	0.32	0.0	0.26	0.1	0.09
Pct. of URM Students – Teacher's School (Per 10%)	-0.3	0.41	0.0	0.26	**-0.9*****	**0.15**
Constant	**81.4*****	**1.47**	**91.8*****	**1.04**	**84.7*****	**0.60**

*** $p < 0.001$ ** $p \leq 0.01$ * $p \leq 0.05$

^ All units are expressed in percentages.

Note: The main predictor variable of interest (i.e., Percent of SE students within a teacher's school) is highlighted in light gray.

Note: Any coefficients that appear to equal zero percent are actually greater than or less than zero percent, but due to rounding rules appear to equal zero percent.

Note: Teacher School Level Taught is omitted from the Year 3 (2015-2016) model as it was not present in the data provided by the NMPED.

Table C14

Model 4b Output, All Years (SPS as Criterion Variable)

	Year 1		Year 2		Year 3	
	2013-2014		2014-2015		2015-2016	
n	2,733		2,738		8,963	
F	5.61		49.25		66.42	
df	14, 2,718		12, 2,725		12, 8,950	
p	0.000		0.000		0.000	
R^2	0.0268		0.1776		0.0746	
	Coef. ^	*SE* ^	Coef. ^	*SE* ^	Coef. ^	*SE* ^
Pct. of ELL Students – Teacher's School (Per 10%)	-0.4	0.33	**0.8*****	**0.16**	**1.9*****	**0.10**
Teacher School Level Taught (Elem. as base)						
Middle	**-2.1****	**0.63**	**-5.7*****	**0.42**	*N/A – Not in*	
High	**-3.9*****	**0.72**	**-8.1*****	**0.50**	*model*	
Teacher URM (No as base)	**1.2***	**0.53**	**1.5*****	**0.37**	**1.7*****	**0.24**
Teacher Years of Experience	**-0.1***	**0.03**	**0.0***	**0.02**	0.0	0.01
Pct. of Gifted Students – Teacher's Classes (Per 10%)	**0.7****	**0.28**	**-1.4*****	**0.25**	**-0.7*****	**0.11**
Pct. of SE Students – Teacher's Classes (Per 10%)	0.0	0.12	**0.3****	**0.09**	**0.1***	**0.05**
Pct. of FRL Students – Teacher's Classes (Per 10%)	0.4	0.30	-0.1	0.25	**0.6*****	**0.09**
Pct. of URM Students – Teacher's Classes (Per 10%)	-0.2	0.34	**-0.8*****	**0.22**	**-0.7*****	**0.13**
Pct. of SE Students – Teacher's School (Per 10%)	**-2.0****	**0.61**	**-1.1****	**0.37**	0.1	0.23
Pct. of FRL Students – Teacher's School (Per 10%)	0.1	0.32	0.0	0.26	0.1	0.09
Pct. of URM Students – Teacher's School (Per 10%)	-0.4	0.41	0.1	0.26	**-0.9*****	**0.15**
Constant	**81.2*****	**1.48**	**91.5*****	**1.04**	**84.8*****	**0.60**

*** $p < 0.001$ ** $p \leq 0.01$ * $p \leq 0.05$

^ All units are expressed in percentages.

Note: The main predictor variable of interest (i.e., Percent of ELL students within a teacher's school) is highlighted in light gray.

Note: Any coefficients that appear to equal zero percent are actually greater than or less than zero percent, but due to rounding rules appear to equal zero percent.

Note: Teacher School Level Taught is omitted from the Year 3 (2015-2016) model as it was not present in the data provided by the NMPED.

Table C15

Model 4c Output, All Years (SPS as Criterion Variable)

	Year 1		Year 2		Year 3	
	2013-2014		2014-2015		2015-2016	
n	2,733		2,738		8,963	
F	2.98		27.09		63.93	
df	12, 2,720		10, 2,727		12, 8,950	
p	0.000		0.000		0.000	
R^2	0.0134		0.0867		0.0814	
	Coef. ^	*SE* ^	Coef. ^	*SE* ^	Coef. ^	*SE* ^
Pct. of FRL Students – Teacher's School (Per 10%)	0.1	0.32	0.1	0.27	-0.1	0.09
Teacher Years of Experience	__-0.1**__	__0.03__	__-0.1**__	__0.02__	0.0	0.01
Pct. of Gifted Students – Teacher's Classes (Per 10%)	__0.8**__	__0.28__	__-1.4***__	__0.24__	__-0.8***__	__0.11__
Pct. of SE Students – Teacher's Classes (Per 10%)	-0.1	0.12	__0.3***__	__0.09__	-0.1	0.05
Pct. of FRL Students – Teacher's Classes (Per 10%)	__0.6*__	__0.30__	0.0	0.27	__0.5***__	__0.09__
Pct. of URM Students – Teacher's Classes (Per 10%)	0.3	0.34	__-0.9***__	__0.23__	__-0.6**__	__0.13__
Student Enrollment – Teacher's School (Per 100 Students)	0.0	0.06	__0.0***__	__0.00__	__-0.3***__	__0.03__
Pct. of SE Students – Teacher's School (Per 10%)	-0.9	0.57	0.1	0.42	-0.1	0.23
Pct. of ELL Students – Teacher's School (Per 10%)	0.2	0.31	__1.9***__	__0.15__	__1.8***__	__0.10__
Pct. of URM Students – Teacher's School (Per 10%)	-0.6	0.40	-0.4	0.28	__-0.6***__	__0.15__
Constant	__79.2***__	__1.53__	__90.3***__	__1.19__	__86.9***__	__0.62__

*** $p < 0.001$ ** $p \le 0.01$ * $p \le 0.05$

^ All units are expressed in percentages.

Note: The main predictor variable of interest (i.e., Percent of FRL students within a teacher's school) is highlighted in light gray.

Note: Any coefficients that appear to equal zero percent are actually greater than or less than zero percent, but due to rounding rules appear to equal zero percent.

Table C16

Model 4d Output, All Years (SPS as Criterion Variable)

	Year 1 2013-2014		Year 2 2014-2015		Year 3 2015-2016	
n	2,733		2,738		8,963	
F	5.62		45.67		63.76	
df	15, 2,717		13, 2,724		13, 8,949	
p	0.000		0.000		0.000	
R^2	0.0302		0.1777		0.0869	
	Coef. ^	*SE* ^	Coef. ^	*SE* ^	Coef. ^	*SE* ^
Pct. of URM Students – Teacher's School (Per 10%)	-0.4	0.41	0.1	0.26	**-0.6*****	**0.15**
Teacher School Level Taught (Elem. as base)					*N/A – Not in model*	
Middle	**-2.3*****	**0.63**	**-5.7*****	**0.42**		
High	**-5.1*****	**0.87**	**-8.3*****	**0.58**		
Teacher URM (No as base)	**1.2***	**0.53**	**1.5*****	**0.37**	**1.7*****	**0.23**
Teacher Years of Experience	**-0.1***	**0.03**	**0.0***	**0.02**	0.0	0.01
Pct. of Gifted Students – Teacher's Classes (Per 10%)	**0.7****	**0.28**	**1.4*****	**0.25**	**-0.8*****	**0.11**
Pct. of SE Students – Teacher's Classes (Per 10%)	0.0	0.12	**0.3****	**0.09**	**0.1***	**0.05**
Pct. of FRL Students – Teacher's Classes (Per 10%)	0.4	0.30	-0.1	0.25	**0.4*****	**0.09**
Pct. of URM Students – Teacher's Classes (Per 10%)	-0.2	0.34	**-0.8*****	**0.22**	**-0.6*****	**0.13**
Student Enrollment – Teacher's School (Per 100 Students)	**0.2****	**0.07**	0.0	0.00	**-0.3*****	**0.03**
Pct. of SE Students – Teacher's School (Per 10%)	**-1.7****	**0.60**	**-1.1****	**0.38**	-0.1	0.23
Pct. of ELL Students – Teacher's School (Per 10%)	-0.4	0.33	**0.7*****	**0.16**	**1.8*****	**0.10**
Pct. of FRL Students – Teacher's School (Per 10%)	0.2	0.32	0.0	0.26	-0.1	0.09
Constant	**79.3*****	**1.58**	**91.2*****	**1.10**	**86.9*****	**0.62**

*** $p < 0.001$ ** $p \leq 0.01$ * $p \leq 0.05$

^ All units are expressed in percentages.

Note: The main predictor variable of interest (i.e., Percent of URM students within a teacher's school) is highlighted in light gray.

Note: Any coefficients that appear to equal zero percent are actually greater than or less than zero percent, but due to rounding rules appear to equal zero percent.

Note: Teacher School Level Taught is omitted from the Year 3 (2015-2016) model as it was not present in the data provided by the NMPED.

APPENDIX D

FACTORS POTENTIALLY AFFECTING TEACHERS' EFFECTIVENESS

This appendix briefly outlines the situational and contextual factors that Kennedy (2010a) discusses in detail as having the potential to affect a teacher's effectiveness. This list is not all-inclusive, so readers are encouraged to review Kennedy's discussion on these factors (see, specifically, pp. 593-596).

1. Parameters of the work itself
    a. Physical space and setup of school and classroom
    b. Time allotments (for teaching, planning, grading, meetings, etc.)
        i. As outlined by the school/district
        ii. As actually utilized
    c. Materials
        i. Curricula frameworks or standards
        ii. Teacher manuals, instructional books, etc.
    d. Work assignments
        i. Number of classes per day/week
        ii. Classes specific to content expertise/credentials
        iii. Additional duties (e.g., extracurricular clubs, recess duty, detention supervision)
2. The students in a classroom and/or school
    a. Inherent immutable traits of students (individually and collectively)
    b. Variable characteristics (e.g., motivation, cooperation, self-reliance, confidence)
    c. Peer effects

3. Institutional practices that interrupt classroom life
    a. Overhead announcements (planned and unplanned)
    b. Administrative interruptions (e.g., telephone calls, office staff visits)
    c. Schedule anomalies (e.g., fire drills, assemblies, field trips, parent-teacher conferences)
4. Policy changes/reforms
    a. Curriculum changes
    b. Class scheduling patterns (e.g., block scheduling)
    c. Student conduct policies (e.g., zero tolerance, conditional zero tolerance; documentation of behavioral incidents)
    d. Administrative changes (e.g., staff meeting days/times, lunch/recess timing)